

**PRV**

PATENT- OCH REGISTRERINGSVERKET  
Patentavdelningen

Rec'd PCT/PTO 15 JUL 2003 / SE 03 / 0 0 0 3 3

20-01-2003

**Intyg  
Certificate**

REC'D 28 JAN 2003

WIPO PCT

Härmed intygas att bifogade kopior överensstämmer med de handlingar som ursprungligen ingivits till Patent- och registreringsverket i nedannämnda ansökan.

This is to certify that the annexed is a true copy of the documents as originally filed with the Patent- and Registration Office in connection with the following patent application.

(71) Sökande Elucidon AB, Lidingö SE  
Applicant (s)

(21) Patentansökningsnummer 0200108-9  
Patent application number

(86) Ingivningsdatum 2002-01-16  
Date of filing

Stockholm, 2003-01-20

För Patent- och registreringsverket  
For the Patent- and Registration Office

*Lina Oljeqvist*  
Lina Oljeqvist

Avgift  
Fee

**PRIORITY DOCUMENT**  
SUBMITTED OR TRANSMITTED IN  
COMPLIANCE WITH  
RULE 17.1(a) OR (b)

PATENT- OCH  
REGISTRERINGSVERKET  
SWEDEN

Postadress/Adress  
Box 5055  
S-102 42 STOCKHOLM

Telefon/Phone  
+46 8 782 25 00  
Vx 08-782 25 00

Telex  
17978  
PATOREG S

Telefax  
+46 8 666 02 86  
08-666 02 86

**BEST AVAILABLE COPY**

jw/

Ref.: 55443 SE

Applicant: Elucidon AB**Information Retrieval****5 THE BACKGROUND OF THE INVENTION AND PRIOR ART**

The present invention relates generally to solutions for information retrieval. More particularly the invention relates to a method of processing digitized textual information according to the preamble of claim 1, a computer program according to claim 10 15, a computer readable medium according to claim 16, a search engine according to the preamble of claim 17, a database according to the preamble of claim 18, a server according to the preamble of claim 20 and a system according to the preamble of claim 21.

15 In this specification, information retrieval is understood as the art of retrieving document related data being relevant to an inquiry from a user. Conventionally, information retrieval systems have been built on the idea that the user actively searches for data by specifying queries (or search phrases) 20 based on keywords (or search terms). Over the past decade, and with the advent of the Internet, the research pertaining information retrieval has grown well past its initial goals of finding methods for efficient indexing and searching.

25 Traditional information retrieval research has been focused on search and retrieval methods based on word indexing and term vector representations. For instance, a vector similarity approach may be used to find relationships and similarities among documents by creating a weighted list of the words (or terms) included in a document. Systems operating according to

this principle can be regarded as "word-comparison" apparatuses, where documents and queries are compared based on the mutual occurrence of words. Nevertheless, if two documents describe the same subject-matter, however with  
5 different words, the method is unable to find a relation between the documents.

To address this problem, and to improve the information retrieval systems, research is currently conducted with the aim at generating conceptual representations of documents. The  
10 conceptual representation involves creating relatively compact term vector representations on basis of a word indexing produced by the earlier known methods. For example, the initial term vectors may be mathematically reduced to a lower dimensionality using a so-called latent semantic indexing.  
15 Another approach is to create a concept representation based on the occurrence of selected concept words. The latter approach is discussed in the master thesis "Artificial Intelligence in an Online Newspaper", Computer Science & Engineering at Linköping Institute of Technology, Sweden, 2000 by Löndahl et  
20 al. and in the international patent application WO00/63837. A feature common to the above methods is that they all result in a document concept distribution, i.e. a weighted list of concept components where the number of concepts is much smaller than the total number of terms. Systems based on such methods may  
25 be used to find relationships between documents, which do not share the same words.

Other examples of research related to the field of the present invention are methods for finding semantic relationships between words. Such relationships are interesting to reveal, for  
30 instance, when performing word disambiguation and when creating thesauruses automatically. Word disambiguation constitutes a considerable challenge in natural language processing and involves deducing the contextual meaning of an ambiguous word, such as "bank", which has a different meaning if the  
35 context is *money* or *river*. Most of the previously proposed

- methods are based on term co-occurrence calculations, i.e. term relationships being calculated based on the frequency at which terms co-occur in the same documents. Research has also been conducted to find a conceptual representation for words based on word proximity in a document corpus. The U.S. patent No. 5,325,298 discloses methods for generating or revising context vectors for a plurality of word stems. The representation thus found may be used to generate the conceptual representation of documents in the document corpus.
- 10 Although, many of today's most advanced information retrieval systems are generally capable of providing an accurate and comparatively relevant search result, there still remains progress to be made in this area. For instance, explicit term-to-term relationships cannot be expressed. Thus, even though
- 15 some of the known methods manage to find documents, which include terms that are synonymous (or by other means equivalent) to a user's search terms, they fail to explain why these documents were encountered. Another problem of the prior-art methods is that the quality of the search result is always limited
- 20 to an upper boundary given by the accuracy of the user's search query. Hence, a poor choice of search phrase inevitably produces a relatively poor search result.

### SUMMARY OF THE INVENTION

- 25 It is therefore an object of the present invention to alleviate the problems above and thus provide an improved solution for processing digitized textual information based on explicit relationships between synonymous terms.

- It is also an object of the invention to offer an information retrieval with an enhanced feedback, which exceeds a maximum result accuracy as given by an initial search phrase.
- 30

According to one aspect of the invention these objects are achieved by a method of processing digitized textual information

as described initially, which is characterized by generating the term-to-concept vectors on basis of the concept vectors. Then, based the term-to-concept vectors for the document corpus, a term-term matrix is generated which describes a term-to-term relationship between the terms in the document corpus. Finally, the processed textual information is derived from the term-term matrix.

An important advantage attained by the term-term matrix is that it provides accurate connections between synonymous terms and related expressions. This in turn, constitutes a basis for accomplishing high quality document searches, i.e. searches in which highly relevant information is identified.

According to a preferred embodiment of this aspect of the invention, each document in the document corpus is associated with a document-concept matrix. The document-concept matrix represents at least one concept element whose relevance with respect to the document is described by a weight factor. The generation of each term-to-concept vector comprises the following steps. First, a term-relevant set of documents is identified in the document corpus. Each document in this term-relevant set contains at least one copy of the term. Second, a term weight is calculated for the term in each of the documents in the term-relevant set. Third, a respective concept vector is retrieved, which is associated with each document in the term-relevant set. However, the term weight must here exceed a first threshold value. Fourth, a relevant set of concept vectors is selected, which includes all concept vectors where at least one concept component exceeds a second threshold value. Fifth, a non-normalized term-to-concept vector is calculated as the sum of all concept vectors in the relevant set. Finally, the non-normalized term-to-concept vector is normalized.

This sub-procedure is advantageous because it accomplishes adequate term-to-concept associations very efficiently. Furthermore, the procedure may be appropriately calibrated with

respect to the application by means of the first and second threshold values.

According to another preferred embodiment of this aspect of the invention, the generation of the term-term matrix comprises the following steps. First, a term-to-concept vector is retrieved for each term in each combination of two unique terms in the document corpus. Second, a relation vector is generated, which describes the relationship between the terms in each combination of two unique terms. Each component in the relation vector is here equal to a lowest component value of the corresponding component values in the term-to-concept vectors. Third, a relationship value is generated for each combination of two unique terms. The relationship value constitutes the sum of all component values in the corresponding relation vector. Finally, a matrix is generated, which contains the relationship values of all combinations of two unique terms in the document corpus.

The term-term matrix per se is a desirable result, since it forms a valuable source of synonymous words and expressions. Furthermore, the above-proposed sub-procedure is attractive because it produces the term-term matrix in a computationally efficient manner.

According to still another preferred embodiment of this aspect of the invention, a statistical co-occurrence value is calculated between each combination of two unique terms in the document corpus. This value describes the dependent probability that a certain second term exists in a document provided that a certain first term exists in the document. The statistical co-occurrence value is then incorporated into the term-term matrix to represent lexical relationships between the terms in the document corpus. The term-term matrix is thus improved by means of a lexical relationship measure, which provides a desirable precision in many applications.

According to yet another preferred embodiment of this aspect of the invention, the processed textual information is displayed on a format, which is adapted for human comprehension, for instance a graphical format. Naturally, such presentation format improves the chances of conveying high-quality information to a user.

According to another preferred embodiment of this aspect of the invention, the displaying step involves presentation of at least one document identifier specifying a document being relevant with respect to at least one term in a query, presentation of at least one term being related to a term in a query, and/or presentation of a conceptual distribution representing a conceptual relationship between two or more terms in the document corpus. The conceptual distribution is based on shared concepts, which are common to said terms.

All these pieces of information represent useful return data and are thus desirable in the information retrieval process.

According to still another preferred embodiment of this aspect of the invention, the displaying step involves presentation of at least one document identifier, which specifies a document being relevant with respect to at least one term in a query in combination with at least one user specified concept. This procedure may include two sub-steps where, in a first step, at least two concepts from the shared concepts in the conceptual distribution are presented to the user. In a second step, the user indicates which concept(s) the query shall be combined with in order to produce a more to-the-point result. This is advantageous since it both vouches for a user-friendly interaction and generates adequate return data.

According to yet another preferred embodiment of this aspect of the invention, the conceptual relationship between a first term and at least one second term is illustrated by means of a respective relevance measure, which is associated with the at

least one second term in respect of the first term. The relevance measure thus indicates the strength of the link between the first and the second term. In most cases this link is asymmetric, i.e. the relevance measure in the opposite direction typically has a different value.

According to another preferred embodiment of this aspect of the invention, the strength in the conceptual relationship between two or more terms is visualized graphically. An advantageous effect thereof is that particular words and expressions being most closely related to each other may be found very efficiently.

According to still another preferred embodiment of this aspect of the invention, the processed textual information is displayed as a distance graph where each term constitutes a node. A node representing a first term is thus connected to one or more other nodes that represent secondary terms to which the first term has a conceptual relationship of at least a specific strength. The relevance measure between the first term and the second term is represented by a least number of node hops there between. This type of distance graph constitutes a first preferred example of a source for deriving a data output in the form of conceptual relationships between words and expressions.

According to another preferred embodiment of this aspect of the invention, the processed textual information is displayed as a distance graph in which each term constitutes a node. A node representing a first term is thus connected to one or more other nodes representing secondary terms to which the first term has a conceptual relationship. Furthermore, each connection is associated with an edge weight, which represents the strength of a conceptual relationship between the terms being associated with the neighboring nodes being linked via the connection in question. The relevance measure between the first term and a particular secondary term is represented by an accumulation of the edge weights being associated with the connections constituting a minimum number node hops between the first



term and the particular secondary term. This type of distance graph constitutes a second preferred example of a source for deriving a data output in the form of conceptual relationships between words and expressions.

- 5 According to yet another preferred embodiment of this aspect of the invention, each term in the document corpus represents either a single word, a proper name, a phrase, or a compound of single words.

- 10 According to another aspect of the invention these objects are achieved by a computer program directly loadable into the internal memory of a digital computer, comprising software for controlling the method described above when said program is run on a computer.

- 15 According to yet another aspect of the invention these objects are achieved by a computer readable medium, having a program recorded thereon, where the program is to make a computer perform the method described above.

- 20 According to still another aspect of the invention these objects are achieved by a search engine as described initially, which is characterized in that the processing unit in turn comprises a processing module and an exploring module. The processing module is adapted to receive the term-to-concept vectors for the document corpus. Based on the term-to-concept vectors, the processing module generates a term-term matrix, which describes a term-to-term relationship between the terms in the document corpus. The exploring module is adapted to receive the query and the term-term matrix. Based on this input, the exploring module processes the term-term matrix and generates the processed textual information.

- 25  
30 This search engine is advantageous, since it is capable of identifying relationships between synonymous words and expression, which typically cannot be found by the prior-art search engines. As further consequence of the proposed search

engine, relevant documents and information can be retrieved that would otherwise have been missed out.

5 According to still another aspect of the invention these objects are achieved by a database as described initially, which is characterized in that it is adapted to deliver the term-to-concept vectors to the proposed search engine. A database where the information has this format is desirable, since it shortens the average response time considerably for a search performed according to the proposed principle.

10 According to a preferred embodiment of this aspect of the invention, the database comprises an iterative term-to-concept engine, which is adapted to receive fresh digitized textual information to be added to the database. Based on the added information, the iterative term-to-concept engine generates  
15 concept vectors for any added document, and generates a term-to-concept vector, which describes a relationship between any added term and each of the concept vectors. An important advantage provided by the iterative term-to-concept engine is that it allows information updates without requiring a complete  
20 rebuilding of the concept vectors and the term-to-concept vectors.

25 According to still another aspect of the invention these objects are achieved by a server as described initially, which is characterized in that it comprises the proposed a search engine, and a communication interface towards the proposed database. This server thus makes searches according to proposed method possible.

30 According to still another aspect of the invention these objects are achieved by a system as described initially, which is characterized in that it comprises the above-proposed server, at least one user client adapted to communicate with the server, and a communication link connecting the at least one user client with the server. Preferably, at least a part of the communication

link is accomplished over an internet (e.g. the public Internet) and the user client comprises a web browser. This browser in turn provides a user input interface via which a user may enter queries to the server. The web browser also receives processed textual information from the server and present it to a user. Hence, a expedient remote access is offered to the information in the database.

Based on an amount of textual data being organized in a document corpus and a method for classifying documents on a conceptual level, the invention thus provides a solution for generating a conceptual representation of all terms in the amount of data on basis of the terms' occurrence in documents and the documents' conceptual classification. A linkage between each term may thereby be expressed by means of a similarity measure. This in turn, is accomplished by identifying the mutual conceptual representations of term combinations followed by a computation of a statistical measure for term co-occurrence. A term-to-term relationship matrix may thus be established. This matrix describes both a conceptual and a lexical similarity between the terms. Moreover, the matrix may be presented graphically, either as a conventional graph or as a relationship network, which is made suitable for human comprehension.

The proposed conceptual representations and relationships allows sophisticated information retrieval operations to be performed, such as finding related terms, identifying subject-matter being common to certain terms and visualizing term relationships. Furthermore, documents being relevant to one or more terms may be retrieved and filtered based on their conceptual representations.

### BRIEF DESCRIPTION OF THE DRAWINGS

The present invention is now to be explained more closely by means of preferred embodiments, which are disclosed as

examples, and with reference to the attached drawings.

- Figure 1 shows a system for providing data processing services according to an embodiment of the invention,
- 5 Figure 2 illustrates, by means of a flow diagram, an indexing pre-processing procedure according to an embodiment of the invention,
- Figure 3 shows a flow diagram, which provides an overview of a method performed by a proposed processing module,
- 10 Figures 4a-c illustrate a sequence according to an embodiment of the invention in which term-to-term relationships are established,
- Figure 5 illustrates, by means of a flow diagram, a method for generating a term-document matrix according to an embodiment of the invention,
- 15 Figure 6 illustrates, by means of a flow diagram, a method for updating a document corpus with added data according to an embodiment of the invention,
- 20 Figures 7a-b illustrate how a term-to-term relationship may be established according to an embodiment of the invention,
- Figure 8 illustrates, by means of a flow diagram, a method for generating a term-term matrix according to an embodiment of the invention,
- 25 Figure 9 illustrates, by means of a flow diagram, the operation of a proposed exploring module according to an embodiment of the invention,
- Figure 10 illustrates, by means of a flow diagram, a method

for finding biased information according to an embodiment of the invention,

- 5      Figure 11      shows an example of a term-term matrix, which is displayed as a relationship network according to an embodiment of the invention,
- Figure 12      shows a flow diagram, which summarizes the proposed method for processing digitized digital information,
- 10      Figure 13      shows a flow diagram, which summarizes a first preferred embodiment of the proposed method for processing digitized digital information, and
- Figure 14      shows a flow diagram, which summarizes a second preferred embodiment of the proposed method for processing digitized digital information.

## 15      DESCRIPTION OF PREFERRED EMBODIMENTS OF THE INVENTION

The following definitions are made with respect to the disclosure of the present invention.

### Document

- 20      Unless otherwise stated, by "document" is meant any textual piece of information written in any language, for instance, an entire text document, a particular part of a document, a document preamble, a paragraph or another sub-part of a text. In addition to the actual information text ("payload") a document
- 25      may include meta information, such as data designating language, author, creation date, images, links, keywords, sounds, video clips etc.

### Proper Name

- 30      Unless otherwise stated, the expression "proper name" is understood as one or more nouns that designate a particular

entity (being or thing). Normally, a "proper name" does not include a limiting modifier and in most English-language cases it is written with initial capital letters. An example of a proper name is "Capitol Hill".

5 Term

Unless otherwise stated, a "term" refers to a single word, a phrase, a proper name, a compound word or another multi-word structure.

Concept

- 10 Unless otherwise stated, by "concept" is meant an abstract or a general idea inferred or derived from specific instances. Usually, a concept may be described by a single word, such as *politics*.

Document Corpus

- 15 Unless otherwise stated, the expression "document corpus" refers to a collection of documents, such as a text archive, a news feed or an article database. A commonly referred document corpus is the Reuters-21578 Text Categorization Test Collection ([www.research.att.com/~lewis/reuters21578.html](http://www.research.att.com/~lewis/reuters21578.html)).

- 20 The present invention relates generally to the field of information retrieval solutions for information exploration. Information exploring here refers to the capability of providing user assistance in extracting specific subsets of information from a larger amount of information. Information exploring also implies finding relations in a given amount of information. According to
- 25 the invention, this can accomplished without the use of Boolean search queries, which is otherwise the standard procedure when working with information retrieval systems.

- 30 The core functionality of the proposed solution is based on a conceptual representation of the terms used in a document corpus and the conceptual relationships between the terms. Based on such relationships, a user can select one or more of the terms and be presented with a related material. The

proposed system is namely capable of presenting related terms, related documents as well as graphical summaries of the selected terms.

Furthermore, by using the generated conceptual relationships, the system is able to graphically display how different pieces of information are related to each other and thereby allow a user to navigate through the information. For example, the relationships between terms can be illustrated by presenting their mutual concepts in a pie chart or by presenting graphical networks of the term relationships. Navigation through the information is enabled by allowing the user to interact with the graphical display of relationships, such as selecting (e.g. by mouse-clicking) a concept in a concept pie chart and thereby exclusively obtain material being related with the selected concept.

Figure 1 shows a system for providing data processing services according to an embodiment of the invention. Digitized textual information, which is presumed to be entered into the system as a document corpus, is stored in a database 130. A server 110 is connected to the database 130 via a communication interface 112. At least one user client 120 may in turn gain access to services provided by the server 110 over a network 140, such as the Internet.

The server 110 contains a search engine 115, which includes a processing unit 150. A processing module 151 in the processing unit 150 transforms the documents (i.e. the digitized textual information in the document corpus of the database 130) into a number of conceptual relationship maps, which describe various relationships in the document corpus.

A user may interact with the system via a user input interface 121a in the user client 120, for example by entering a query Q. The query Q is forwarded to the server 110 over a first communication link 141 and an interface 116. Based on the

user's interaction with the system, for instance, choosing a certain term in the query Q, an exploring module 152 extracts relevant processed textual information R, which is produced from the relationships generated by the processing unit 150. The processed textual information R is then returned to the user client 120 via a second communication link 142 and presented to the user via a user output interface 121b. Preferably, the information R is displayed in a graphical format that allows further interaction with the information R.

Figure 2 illustrates, by means of a flow diagram, an indexing pre-processing procedure according to an embodiment of the invention. This procedure may be performed by a proposed indexing engine 320, which will be described further below with reference to the figures 3 and 6. The pre-processing involves extracting all terms included in an unformatted text and assigning weights to each of the terms based on their information content. A list of terms and a term-document matrix (TDM) are generated as a result of this indexing.

The TDM is a  $N \times M$  matrix containing M vectors of dimensionality N, where N represents the number of unique terms in the document corpus (usually approximately equal to the number of words in the language of the document corpus) and M represents the number of documents in the corpus. Each vector component in the TDM contains a weight in the interval [0,1], which indicates the importance of a term in a document, or vice versa, i.e. the importance of a document to a given term.

The indexing pre-processing procedure includes the following steps. A first step 210 performs word splitting. This means that the text is split into a number of words, based on an "allowed character" rule. The definition of what is an "allowed character" depends on the language. Usually, at least all characters included in the language's alphabet(s) are allowed. A character in the text, which is however not contained in the set of allowed characters results in a word split. Typically, a word splitting is



performed when a space-character occurs.

Subsequently, a step 220 performs proper name identification in the text. The step 220 thus identifies compound terms consisting of two or more terms, such as "Bill Clinton". The treatment of  
5 each proper name as a single term lowers the error rate in the information retrieval process, since ambiguities are thereby reduced. An example of an ambiguity that will occur unless proper name identification is performed is that between "Carl  
10 Lewis" and "Lennox Lewis". Here, the term "Lewis" would erroneously cause the search engine 115 to judge a document containing "Carl Lewis" and another containing "Lennox Lewis" to be related to each other.

After that, a step 230 removes any stop words in the text. Some  
15 terms namely have a low or no importance to the content of a text. Preferably, such insignificant terms are removed according to a language-specific stop word list. The words "the", "a", "is" and "are" in the English language constitute typical examples of stop words to be removed.

Then, a step 240 applies a stemming algorithm. This algorithm  
20 ensures that different forms of words that have the same word stem are treated as a single term. Naturally, the stemming algorithm must be language-specific and it is applied to all words in the text. The algorithm removes any word suffixes and transforms the words into their common word stem. A commonly  
25 used algorithm for stemming in an English text is the Porter stemming algorithm. Based the principles behind this algorithm, the person skilled in the art may design a stemming algorithms for any other language.

Following the step 240, a step 250 performs term weighting of  
30 the words in the text. Thereby, each unique term in each document is assigned a weight according to its information content. The so-called Term Frequency times Inverse Document Frequency (TFIDF) is a commonly used method for this.

According to a preferred embodiment of the invention, the information content in a document is determined by using an extension to the traditional TFIDF term weighting scheme. Specifically, a term position parameter  $p(t,d)$  (which will be explained below) is added to each term.

A certain term  $t$  in a document  $d$  is thus allocated a weight  $w(t, d)$  in a document  $d$  according to:

$$w(t,d) = \frac{n(t,d)}{n(d)} \cdot -\log\left(\frac{N(t,D)}{N(D)}\right) \cdot p(t,d)$$

where  $n(t,d)$  is the number of occurrences of the term  $t$  in the document  $d$ ,

$n(d)$  is the total number of terms in the document  $d$ ,

$N(t,D)$  is the number of documents in which the term  $t$  exists,

$N(D)$  is the total number of documents in the document corpus, and

$p(t,d)$  is a domain specific weight function dependent on the positions of the term  $t$  in the document  $d$ .

The parameter  $p(t, d)$  is used to increase the importance of a term occurring in, for instance, the title or preamble of a document. For example, a term occurring in the headline may have  $p(t,d) = 3.0$ , while it has  $p(t,d) = 1.0$  when occurring in the body text.

Finally, a step 260 normalizes the vectors in the TDM. Preferably, the normalization is performed according to the Euclidean norm. Thus, for a term  $t_i$  in a document  $d_k$  (i.e. position  $(i,k)$  in the term-document matrix) the normalization  $w(t_i, d_k)$  is given by:

$$w(t_i, d_k) = \frac{w(t_i, d_k)}{\sqrt{\sum_{j=1}^N w(t_j, d_k)^2}}.$$

Figure 3 shows a flow diagram, which provides an overview of the method performed by the processing module 151 in figure 1. The processing module 151 performs a number of processing steps and calculations in order to generate relationship matrices that describe various relationships within the document corpus. In this context, a relationship is indicated by a numeric value, which describes for example the similarity between two terms in the document corpus. The figure shows a set of engines 320, 340, 360 and 380 and illustrates how these together process the various data types according to the invention.

A document corpus 310 containing at least one document is presumed to be entered on a digital format and there after be stored in a computer memory storage system, such as the database 130 in figure 1. An indexing engine 320 extracts every term found in the document corpus 310, preferably according to the indexing pre-processing procedure described with reference to figure 2 above. The indexing engine 320 also assigns weights to the extracted terms (step 250 in figure 2), which specifies the terms' information importance relative to the document in which they occur.

A document-concept-matrix (DCM) 390 describes how the documents in the document corpus 310 are related to concepts. Each document in the corpus 310 is thereby described by a normalized vector in the DCM 390, which denotes a distribution of concepts describing the particular document. For instance, in a news domain a document titled "Tony Blair attempts to save the peace-process in Northern Ireland" would typically have a concept distribution that indicates high relationships to the concepts "UK", "Northern Ireland", "Negotiations" and "Government".

A term-document matrix (TDM) 330 describes how terms occur in documents. Each unique term in the document corpus 310 has a normalized vector in the TDM 330, which denotes a distribution of documents that contain the term and the term's importance in these documents. In the art of information retrieval this matrix is commonly referred to as an inverted index.

A term-concept matrix engine 340 receives the DCM 390 and the TDM 330, and on basis thereof generates a matrix of vectors, which contains weight values representing relationships between terms and concepts. In the DCM 390, each document is associated with a concept vector via different weight values, and in the TDM 330, each term has a weighted value with respect to each document vector in which it occurs.

The matrix produced by the engine 340 is an  $N \times M$  dimensional array of normalized term vectors, which each contains a set of weight values.  $N$  here represents the number of unique terms in the document corpus and  $M$  represents the number of concepts.

The weight value lies in the interval  $[0,1]$  and indicates how closely a term is associated with a particular concept, based on the context in which the term has appeared. A high weight thus indicates a close relationship. For example, the term "NHL" is likely to have a high relationship with the concept "Hockey". The procedure according to which the term-to-concept relationships are generated will be further illustrated below with reference to the figures 4a-c.

A term-concept matrix (TCM) 350 describes how the terms are related to concepts. Each unique term in the corpus 310 has a normalized vector in the TCM 350, which denotes a distribution of concepts describing the document. For instance, in a news domain the term "Bill Clinton" would typically have a concept distribution indicating the concepts "President", "Government" and "US".

A term-term matrix engine 360 receives the TDM 330 and the TCM 350, and on basis thereof generates a term-term matrix 370, which contains vectors that describe conceptual relationships between the terms.

- 5 The term-term matrix (TTM) 370 describes how each term is related to each of the other terms in the corpus 310. Hence, each unique term in the corpus 310 has an entry in the TTM 370, which denotes a distribution vector of terms being related to the term. For instance, in a news domain the term "Bill Clinton" would typically have a term distribution including "George Bush", "Al Gore" and "Hillary Clinton".

15 A document-concept matrix engine 380 is used to generate conceptual representations of any new documents being entered into the system, either at system start-up when a complete document corpus 310 is entered or when updating the corpus 310 with one or more added documents. A preferred procedure for accomplishing such information update is described below and with further reference to figure 6. However, any alternative method known from the prior art may equally well be used. In any case, the engine 380 updates the DCM 390 based on the TDM 330 and the TCM 350.

25 The document-concept matrix engine 380 produces the conceptual distribution for a document, i.e. a description of the relationships between the document and all concepts in the corpus 310. In essence, the documents are processed by means of algorithms that find a conceptual document description. This description has the property that documents, which relate to the same topics, or basically has the same semantic meaning, will receive a similar conceptual description. Any of the prior-art methods for generating conceptual descriptions of documents may be used for this provided that the result thereof can be expressed as a DCM, where each row is a normalized document vector, which denotes a distribution of concepts describing each document in the document corpus 310.

Formally the engine 380 calculates, for each document  $D_i$  and concept  $C_j$ , a document-concept relationship value  $rdc(D_i, C_j)$  according to:

$$rdc(D_i, C_j) = \frac{rdc(D_i, C_j)}{\sqrt{\sum_{i=1}^M rdc(D_i, C_j)^2}}$$

- 5 and forms a matrix of the relationship value  $rdc(D_i, C_j)$  as elements, where each element  $(i, j)$  in the matrix contains the row-wise normalized  $rdc(D_i, C_j)$  value.

Due to the normalization, the range of the  $rdc(D, C)$  is  $[0, 1]$ . A value close to 1 thus indicates a close conceptual relationship  
 10 between the document and a concept, while a value close to 0 indicates no or an insignificant relationship.

Figures 4a-c illustrate a sequence according to an embodiment of the invention in which term-to-term relationships are established. A set of documents 411 - 414 in a document  
 15 corpus are presumed to be related to a number of concepts 420 - 424 as illustrated by the arrows. Furthermore, a first term 431 ("Carl Bildt") and a second term 432 ("Tony Blair") are weighted in all documents 411, 412 in which they occur (see figure 4b). Based on the fact that terms 431, 432 are related to the  
 20 documents 411; 412 and documents 411; 412 in turn are related to the concepts 421 - 423, the term-concept matrix engine (340 in figure 3) is able to compute term-to-concept relationships between the first term 431 ("Carl Bildt") and a second concept 422 ("Kosovo") as shown in figure 4c.

25 In this example, the first term 431 ("Carl Bildt") occurs in a first document 411 and in a second document 412. The first document 411 is in turn related to a first concept 421 ("Kosovo") and the second concept 422 ("UN"), while the second document 412 is only related to the second concept 422 ("UN"). Thus, the  
 30 first term 431 ("Carl Bildt") is related to both the first concept 421 ("Kosovo") and to the second concept 422 ("UN"), however,

the relationship to the second concept 422 ("UN") being stronger.

A more exact description of this algorithm is described below with reference to figure 5. Here, a flowchart illustrates the different operations performed by the term-concept matrix engine (340 in figure 3) and how they interact with each other. Based on the DCM 390, the processing starts in a step 510 by iterating over all unique terms in the document corpus (310 in figure 3). A step 520, performs, for each term  $t_i$ , a second iteration over all concepts. The algorithm thus traverses over all positions in the resulting TCM (350 in figure 3). A step 530 calculates a relation value  $rtc(t_i, c_j)$  for a given term  $t_i$  and a given concept  $c_j$ , according to:

$$rtc(t_i, c_j) = \sum_{\{k | t_i \in d_k\}} w(t_i, d_k) \cdot rdc(d_k, c_j).$$

- 15 The sum is computed over all documents containing a term  $t_i$ . The factor  $w(t_i, d_k)$  represents a weighted value for the term  $t_i$  in a document  $d_k$  as computed by the indexing engine (320 in figure 3). The factor  $rdc(d_k, c_j)$  is a value that describes a relationship between the document  $d_k$  and the concept  $c_j$  as
- 20 specified in the DCM (390 in figure 3). According to a preferred embodiment of the invention, all documents having a  $w(t_i, d_k)$ -value below a first threshold (see step 1330 in figure 13) and each document having all its  $rdc(d_k, c_j)$ -values below a second threshold (see step 1340 in figure 13) are ignored. This namely
- 25 reduces the noise and thus ensures that a term's conceptual representation is exclusively based on those documents where the term has a particular significance, and where the documents in turn can be described by a comparatively distinct conceptual representation.
- 30 The resulting sum represents a weighted relationship between a certain term and a certain concept. In a step 540, the sum is normalized  $rtc$  using Euclidean norm:

$$rtc(t_i, c_j) = \frac{rtc(t_i, c_j)}{\sqrt{\sum_{j=1}^M rtc(t_i, c_j)^2}}$$

The normalized rtc-values for a specific term are stored in the TCM (350 in figure 3) at their respective positions (i, j), thus forming a normalized term-to-concept row-vector at row i. The document-concept engine 380 iteratively updates the DCM (390 in figure 3) accordingly.

Figure 6 illustrates, by means of a flow diagram, a method for updating a document corpus with added data according to an embodiment of the invention. When the TCM 350 has been generated, it can be used to iteratively assign a conceptual distribution to new, previously unknown terms appearing in an added document.

In a first step 610, a document  $d_k$  enters the indexing engine 320 where it is processed. For terms  $t_i$  (where  $i = 1, \dots, m$ ) with an existing conceptual distribution, a step 620 retrieves the distribution row vector from the TCM 350. The step 620 also retrieves a corresponding weight value for the term  $t_i$  in the document  $d_k$  from the TDM 330.

A step 650 calculates term-to-concept vectors for each added and previously unknown term  $t_j$  (where  $j = m+1, \dots, n$ ) by iterating over all concepts (step 640), for each concept  $c_s$ , its cumulative weight  $rtc(t_{new}, c_s)$  in the document  $d_k$  according to:

$$rtc(t_{new}, c_s) = \sum_{i=1}^m rtc(t_i, c_s) \cdot rtd(t_i, d_k)$$

A step 670 then assigns the cumulative weight  $rtc(t_{new}, c_s)$  for the concept  $c_s$  to each of the previously unclassified terms (step 660) in the added document  $d_k$ .

The term-to-concept relationship values for the added terms  $t_j$  are finally normalized using Euclidean norm in a step 680. The



normalized rtc-values for term  $t_j$  are stored in the TCM 350 at their respective positions  $(j, s)$ , thus forming a normalized term-to-concept row-vector at row  $j$ .

5 The term-term matrix engine (360 in figure 3) generates an  $N \times N$  relationship matrix of all terms in the document corpus, where  $N$  is the number of unique terms in the corpus. A relationship value in the interval  $[0, 1]$  is generated from each term to every other term. The generation of the term-term matrix uses the TCM in conjunction with a term co-occurrence calculation, which is  
10 described below with reference to figures 7a-b. The merit of combining the two methods is that both conceptual and lexical similarities can thereby be described with a single similarity measure.

15 The idea of using the TCM (which may also be regarded as a network, see figure 11) in order to find relationships between terms will now be elucidated. Based on relationships between a set of terms 431 - 434 and a set of concepts 420 - 424, term-to-term relationships can be generated by identifying mutual, or shared, concept components. As an example, a first term 431  
20 ("Carl Bildt") and a sixth term 436 ("Bill Clinton") would be conceptually related, since they are both related to a first concept 421 ("Kosovo") and a second concept 422 ("UN"), see bold lines figure 7b.

25 Figure 8 illustrates, by means of a flow diagram, a method for generating a term-term matrix according to an embodiment of the invention. Two initial steps 810 and 820 in combination with two loop-back steps 841 and 861 respectively accomplish a double iteration over all unique terms  $t_i \neq t_j$  in the document corpus. Thereby, a relation value is generated which describes  
30 the relationships between any specific term and each of the other terms.

For each pair of terms  $t_i$  and  $t_j$ , a step 830 calculates a  $rttc(t_i, t_j)$ -value as the sum of the lowest term-concept relationship values

over all concepts. This corresponds to the expression:

$$rtc(t_i, t_j) = \sum_{k=1}^m \min(rtc(t_i, c_k), rtc(t_j, c_k))$$

where  $c_k$  specifies a certain concept,

$m$  represents the total number of concepts, and

5  $rtc(t, c)$  is the relationship value defined in the TCM as described above.

10 The minimum-function produces the effect that the conceptual relationships are here defined by the mutual concepts for the terms. All the iterations (steps 810 and 820) result in a description of the conceptual relationships between all terms in the form of a primary term-to-term matrix.

15 In order to improve the precision of this matrix, the relationship values between terms are enhanced in a step 840 based on their statistical co-occurrence in the document corpus. Two terms are defined as co-occurring if they are found in the same document(s). A co-occurrence value  $rtto(t_i, t_j)$  is generated, based on the dependent probability  $p(t_i \in d_k | t_j \in d_k)$  that a certain term  $t_j$  exists in a document  $d_k$  chosen at random, provided that  $t_i$  exists in  $d_k$ . This definition is equivalent to the  
20 expression:

$$rtto(t_i, t_j) = p(t_i | t_j) = \frac{p(t_i \cap t_j)}{p(t_j)}$$

25 The probabilities above are easily calculated using the TCM. For example, in a certain document corpus the term "NHL" and the term "hockey" may co-occur in 5% of the documents. In the same corpus, the term "NHL" is presumed to occur in 10% of the documents. The dependent probability of finding the term "hockey" given the term "NHL" is thus  $0.05/0.10=0.5$ . In other words, the co-occurrence between "NHL" and "hockey", i.e. the  $rtto$ -value, is  $rtto(\text{"hockey"}, \text{"NHL"}) = 0.5$ .

In a step 850, the two term-term relationship metrics are then combined into a final term-term relationship value  $rtt$ , which replaces the initial  $rttc$ -value in the primary term-to-term relationship matrix according to:

$$5 \quad rtt(t_i, t_j) = \alpha \cdot rtto(t_i, t_j) + \beta \cdot rttc(t_i, t_j)$$

where  $\alpha$  and  $\beta$  represent a first and a second constant, which define the importance of the  $rttc$ - and  $rtto$ -values respectively. The choice of  $\alpha$  and  $\beta$  thus controls the influence of conceptual and lexical relationships in the final term similarity measure.

- 10 Both the constants  $\alpha$  and  $\beta$  may be chosen arbitrarily, since the  $rtt$ -values are normalized using Euclidean norm in a following step 860. The matrix is normalized row-wise for a row  $i$  as follows:

$$rtt(t_i, t_j) = \frac{rtt(t_i, t_j)}{\sqrt{\sum_{j=1}^N rtt(t_i, t_j)^2}}$$

- 15 where  $N$  is the total number of terms unique terms in the document corpus. As a result, the term-term matrix 375 is produced.

- 20 Please note that the co-occurrence value is based on a non-symmetric function, i.e. typically  $rtto(t_i, t_j) \neq rtto(t_j, t_i)$ . In most cases, the term-term relationship matrix is hence non-symmetric. This, corresponds to the case where a first term has a strong relationship to a second term, without however the second term having a strong relationship to the first term. For example, the term "Mike Tyson" may have a very strong relationship to the term "boxing" whilst the term "boxing" only is weakly related to the term "Mike Tyson".

25 Figure 9 illustrates, by means of a flow diagram, the operation of the exploring module (152 in figure 1) according to an embodiment of the invention. The exploring module is used to provide

services based on relationships in the document corpus. Based on one or a plurality of terms, the module then presents relevant documents, related terms and a conceptual distribution.

5 A joint concept engine (JCE) 920 is here used to determine the concepts being common to at least two terms 910. The terms 910 are input to the TCM 350 and the concept distribution for each term (corresponding to the respective term's row in the TCM) is sent as input to the JCE 920. The JCE 920 calculates a joint concept distribution by selecting the lowest component  
10 values from all the terms' concept vectors, which are given by the TCM 350. A new vector is created based on these component values. The vector is subsequently normalized and returned as the result from the JCE 920. The result from the JCE 920 may be regarded as an explanation of the conceptual  
15 relationship between two or more terms. For example, a user asking for the joint concepts pertaining to the terms "Madeleine Albright" and "Tony Blair" may be presented with a piechart covering the concepts "Politics" and "Balkan War".

20 A concept bias engine (CBE) 940 is used to retrieve a set of relevant documents, given at least one term, which not only relates to the given term(s), however also relates to at least one concept. The latter may be supplied directly from a user, from a subsystem or a search engine in a step 935. For example, the at least one concept may be selected from all concepts occurring  
25 in the term's conceptual distribution, such that information will be retrieved that is related to the term in a specific way.

30 If no concept is used as input to the CBE 940 via the step 935, the result will be a set of documents 945 being related to the given term(s) 910 without any bias. However, if a concept distribution is input to the CBE 940 in the step 935 this will "bias" the set of documents 945, or re-arrange this set, based on the documents' 945 proximity to the given distribution. Specifically, the biasing is produced on basis of the documents' conceptual representation as given by the DCM 390.

Returning to the example stated above, a further illustrating example is here presented in order to illustrate the use of the CBE 940. A user who selects the term "Madeleine Albright" would initially be presented with related terms, related documents, and say, a piechart including the concepts "Politics",  
 5 "Balcan War" and "America". If the user subsequently selects the concept "Balcan War", the CBE 940 will present documents that not only relates to "Madeleine Albright", however specifically concerns the "Balcan War". Thus, the user is guided  
 10 into finding specific subsets of the document corpus that may be of particular interest to him/her.

Figure 10 illustrates, by means of a flow diagram, a method for finding biased information according to an embodiment of the invention. Based on a set of selected terms  $T_1, \dots, T_n$  being  
 15 entered in a first step 1010, a following step 1020 uses the TDM to find documents  $D_i$  that contain these terms  $T_1, \dots, T_n$ .

Given the documents'  $D_i$  conceptual distributions  $C_j$ , as indicated by the DCM in a step 1030, and an input bias conceptual distribution  $B_{CD}$  received via a step 1050 in a step  
 20 1040, a step 1060 calculates a relationship value  $rcc(C_j, B_{CD})$  for each document  $D_i$  according to:

$$rcc(C_i, B_{CD}) = \sum_{k=1}^n C_{i,k} B_{CD,k}$$

where  $C_{i,k}$  is a weight for a concept  $k$  in the distribution  $C_i$  and  $B_{CD,k}$  is a weight for the concept  $k$  in the distribution  $B_{CD}$ . The  
 25 sum is calculated over every concept. If the concept distributions  $C_i$  are represented as vectors, the  $rcc$ -function is equivalent to the so-called dot product. Finally, resulting documents are returned in a step 1070. These documents are ranked in descending order by the value in the  $rcc$ -function.

30 Please observe the loop from the step 1010, via the step 1050 to the step 1040. According to a preferred embodiment of the invention, based on the term input and the JCE (920 in figure 9),

a number of concepts suitable for biasing are presented to the user.

Returning now to figure 9. The purpose of the path engine 960 is to describe relationships between terms by using the term-term matrix 370 plus at least one term as the input. The path engine 960 has two modes of operation, Single Term Mode (STM) and Multiple Terms Mode (MTM).

In STM, one and only one term is supplied as input. The primary purpose of STM is to find the most relevant terms for a specific term. For example, if "Yasser Arafat" were used as input, the path engine 960 would typically reply "Israel", "Benjamin Netanyahu" and "Bill Clinton" as well as corresponding relevance measures for each term. The path engine 960 uses the term-term matrix 370 as a graph matrix, and traverses this graph to find any terms being related to the input. All terms within a certain distance in the graph are then returned as a result from the engine 960. The distance measure may differ depending on implementation, however reasonable measures are either the number of graph nodes from input or the accumulated edge weights in the graph.

In MTM, a plurality of terms are instead supplied as input. The path engine 960 again uses the term-term matrix 370 as a graph matrix, and uses well-known graph algorithms to calculate and return a sub-graph of this graph. As in STM, the algorithms apply a distance measure that depends on the specific implementation. The same distance measures as above may be applied. The choice of graph algorithm determines the use of the sub-graph. For instance, Dijkstra's Shortest Path algorithm provides the shortest path between two terms in the graph. Floyd-Warshall's algorithm provides the shortest paths between all supplied terms. The so-called MST provides the minimal spanning tree spanning all supplied terms. The purpose of the various sub-graphs is to examine the relationship between a plurality of terms, and to allow the relationship to be graphically

visualized to enable users to further explore the information in the system.

5 An example of the use of MTM is shown in Figure 11. The figure shows a term-term matrix being displayed as a relationship network. Here, a first term 431 ("Carl Bildt"), a second term 433 ("Gerhard Schröder") and a third term 434 ("Hillary Clinton") are presumed to be used as input to a path engine 960 running in MTM mode, with Floyd-Warshall as the chosen algorithm and number of graph nodes from input as the distance measure. The  
10 path engine 960 calculates the shortest distance between all three terms 431, 433 and 434 in the graph. These paths are displayed as dashed lines in the figure.

As is apparent from the figure, there are three equidistant relationship paths between the first term 431 "Carl Bildt" and the  
15 second term 433 ("Gerhard Schröder"). These paths run via a fourth term 432 ("Tony Blair"), a fifth term 435 ("Kofi Annan") and a sixth term 436 ("Bill Clinton") respectively.

Furthermore, the shortest possible path from the first term 431 ("Carl Bildt") and the second term 433 ("Gerhard Schröder") to a  
20 seventh term 434 ("Hillary Clinton") run via the sixth term 436 ("Bill Clinton"). The merit of the MTM is that it reveals implicit relations between terms, such as "proper names". Moreover, the relationships may easily be explained and displayed graphically to a user, thus allowing him/her to further explore the information in search of relevant facts.  
25

In order to sum up, the general method for processing digitized digital information according to the invention will now be described with reference to figure 12. The information is presumed to be organized in terms, documents and document  
30 corpora, where each document contains at least one term and each document corpus contains at least one document.

A first step 1210 generates a concept vector for each document in a document corpus. The concept vector conceptually

classifies the contents of the document on a relatively compact format. A following step 1220 generates, for each term in the document corpus, a term-to-concept vector which describes a relationship between the term and each of the concept vectors.

- 5 Subsequently, a step 1230 generates a term-term matrix, which describes a term-to-term relationship between the terms in the document corpus. The term-term matrix is produced on basis of the term-to-concept vectors for the document corpus. Finally, a  
10 step 1240 processes the term-term matrix into processed textual information, which preferably has a graphical format that is well adapted to be comprehended by a human user.

- Figure 13 shows a flow diagram, which summarizes a sub-procedure for generating a term-to-concept vector according to a preferred embodiment of the invention. Each document in the  
15 document corpus is here presumed to be associated with a document-concept matrix, which represents at least one concept element whose relevance with respect to the document is described by a weight factor.

- A first step 1310 identifies a term-relevant set of documents in  
20 the document corpus. Each document in the term-relevant set contains at least one occurrence of the term. Then, a step 1320 calculates a term weight for the term in each of the documents in the term-relevant set. A step 1330 there after, retrieves a respective concept vector being associated with each document  
25 in the term-relevant set. However, a condition for including a specific concept vector is that the term weight therein exceeds a first threshold value. Subsequently, a step 1340 selects a relevant set of concept vectors including any concept vector in which at least one concept component exceeds a second  
30 threshold value. A step 1350 then calculates an initial non-normalized term-to-concept vector as the sum of all concept vectors in the relevant set. Finally, a step 1350 normalizes the initial term-to-concept vector that was obtained in the step 1350. Preferably, the normalizing is carried out according to the  
35 Euclidian norm.



Figure 14 shows a flow diagram, which summarizes a sub-procedure for generating the term-term matrix according to a preferred embodiment of the invention. A first step 1410 retrieves a respective term-to-concept vector for each term in  
5 each combination of two unique terms in the document corpus. Then, a step 1420 generates a relation vector, which describes the relationship between the terms in each combination of two unique terms. Each component in the relation vector is here equal to a lowest component value of corresponding component  
10 values in the term-to-concept vectors. A subsequent step 1430, generates a relationship value for each combination of two unique terms as the sum of all component values in the corresponding relation vector. Finally, a step 1440 generates a matrix, which contains the relationship values of each combination of two unique terms in the document corpus.  
15

All of the process steps, as well as any sub-sequence of steps, described with reference to the figures 12 - 14 above may be controlled by means of a computer program being directly loadable into the internal memory of a computer, which includes  
20 appropriate software for controlling the necessary steps when the program is run on a computer. Naturally, the same is also true with respect to the procedures described with reference to the figures 2 - 11. Furthermore, such computer programs can be recorded onto arbitrary kind of computer readable medium as  
25 well as be transmitted over arbitrary type of network and transmission medium.

The term "comprises/comprising" when used in this specification is taken to specify the presence of stated features, integers, steps or components. However, the term does not preclude the  
30 presence or addition of one or more additional features, integers, steps or components or groups thereof.

The invention is not restricted to the described embodiments in the figures, but may be varied freely within the scope of the claims.

### Claims

1. A method of processing digitized textual information, the information being organized in terms, documents and document corpora, where each document contains at least one term and each document corpus contains at least one document, the method comprising:
- 5 generating a concept vector for each document in a document corpus, the concept vector conceptually classifying the contents of the document on a relatively compact format, and
- 10 generating, for each term in the document corpus, a term-to-concept vector describing a relationship between the term and each of the concept vectors, characterized by the term-to-concept vectors being generated on basis of the concept vectors, and the method comprising:
- 15 receiving the term-to-concept vectors for the document corpus and on basis thereof generating a term-term matrix describing a term-to-term relationship between the terms in the document corpus, and
- 20 processing the term-term matrix into processed textual information.
2. A method according to claim 1, characterized by each document in the document corpus being associated with a document-concept matrix representing at least one concept element whose relevance with respect to the document is described by a weight factor, the generation of each term-to-concept vector comprising:
- 25 identifying a term-relevant set of documents in the document corpus, each document in the term-relevant set containing at least one occurrence of the term,
- 30 calculating a term weight for the term in each of the documents in the term-relevant set,
- retrieving a respective concept vector being associated with each document in the term-relevant set where the term weight exceeds a first threshold value,

selecting a relevant set of concept vectors including any concept vector in which at least one concept component exceeds a second threshold value,

- 5       calculating a non-normalized term-to-concept vector as the sum of all concept vectors in the relevant set, and  
normalizing the non-normalized term-to-concept vector.

3. A method according to any one of the preceding claims, **characterized by** the generation of the term-term matrix comprising:

- 10       retrieving, for each term in each combination of two unique terms in the document corpus, a respective term-to-concept vector,

- generating a relation vector describing the relationship between the terms in each combination of two unique terms, each component in the relation vector being equal to a lowest component value of corresponding component values in the term-to-concept vectors,

- 15       generating a relationship value for each combination of two unique terms as the sum of all component values in the corresponding relation vector, and

- 20       generating a matrix containing the relationship values of all combinations of two unique terms in the document corpus.

4. A method according to any one of the preceding claims, **characterized by**

- 25       calculating a statistical co-occurrence value between each combination of two unique terms in the document corpus, the statistical co-occurrence value describing a dependent probability that a certain second term exists in a document provided that a certain first term exists in the document, and

- 30       incorporating the statistical co-occurrence values into the term-term matrix to represent lexical relationships between the terms in the document corpus.

5. A method according to any one of the preceding claims, **characterized by** displaying the processed textual information on a format being adapted for human comprehension.
- 5 6. A method according to claim 5, **characterized by** the displaying step involving presentation of at least one of:  
at least one document identifier specifying a document being relevant with respect at least one term in a query,  
at least one term being related to a term in a query, and  
10 a conceptual distribution representing a conceptual relationship between two or more terms in the document corpus, the conceptual distribution being based on shared concepts which are common to said terms.
- 15 7. A method according to claim 6, **characterized by** the displaying step involving presentation of at least one document identifier specifying a document being relevant with respect to at least one term in a query in combination with at least one user specified concept.
- 20 8. A method according to claim 7, **characterized by** selecting the at least one user specified concept from the shared concepts in the conceptual distribution.
- 25 9. A method according to any one of the claims 5 - 8, **characterized by** illustrating the conceptual relationship between a first term and at least one second term by means of a respective relevance measure being associated with the at least one second term in respect of the first term.
10. A method according to claim 9, **characterized by** displaying the processed textual information on a graphical format which visualizes the strength in the conceptual

relationship between at least two terms.

11. A method according to any one of the claims 9 or 10, characterized by displaying the processed textual information as a distance graph in which

- 5 each term constitutes a node,  
a node representing a first term is connected to one or more other nodes representing secondary terms to which the first term has a conceptual relationship of at least a specific strength, and
- 10 the relevance measure between the first term and the at least one second term is represented by a minimum number of node hops between the first term and the at least one second term.

12. A method according to any one of the claims 9 or 10, characterized by displaying the processed textual information as a distance graph in which

- 15 each term constitutes a node,  
a node representing a first term is connected to one or more other nodes representing secondary terms to which the first term has a conceptual relationship, each connection is associated with an edge weight representing the strength of a conceptual relationship between the first term and a particular secondary term, and
- 20 the relevance measure between the first term and a particular secondary term is represented by an accumulation of the edge weights being associated with the connections constituting a minimum number node hops between the first term and the particular secondary term.
- 25

13. A method according to any one of the preceding claims, characterized by each term representing one of:  
a single word,

30

a proper name,  
a phrase, and  
a compound of single words.

14. A method according to any one of the preceding claims,  
5 characterized by updating the document corpus with added data in form of at least one new document by means of:  
identifying any added terms in the new document which lack a representation in the document corpus,  
identifying any existing terms in the new document which  
10 were represented in the document corpus before adding the at least one new document,  
retrieving, for each of the existing terms, a corresponding concept vector,  
generating a new concept vector with respect to the at  
15 least one new document as a sum of the corresponding concept vectors,  
normalizing the new concept vector into a normalized new concept vector, and  
assigning the normalized new concept vector to each of  
20 the added terms in the new document.

15. A computer program directly loadable into the internal memory of a digital computer, comprising software for performing the steps of any of the claims 1 – 14 when said program is run on a computer.

- 25 16. A computer readable medium, having a program recorded thereon, where the program is to make a computer perform the steps of any of the claims 1 – 14.

17. A search engine (115) for processing an amount of digitized textual information and extracting data there from, the

information being organized in terms, documents and document corpora, where each document contains at least one term and each document corpus contains at least one document, comprising:

- 5       an interface (116) adapted to receive a query (Q) from a user, and  
      a processing unit (150) adapted to process a document corpus on basis of the query (Q) and return processed textual information (R) being relevant to the query (Q), said process  
10   involving  
      generating a concept vector for each document in the document corpus, the concept vector conceptually classifying the contents of the document on a relatively compact format, and  
15       generating, for each term in the document corpus, a term-to-concept vector describing a relationship between the term and each of the concept vectors,  
      characterized in that the processing unit (150) in turn comprises:  
      a processing module (151) adapted to receive the term-to-concept vectors for the document corpus and on basis thereof  
20   generate a term-term matrix describing a term-to-term relationship between the terms in the document corpus, and  
      an exploring module (152) adapted to receive the query (Q) and the term-term matrix, and on basis of the query (Q),  
25   process the term-term matrix into the processed textual information (R).

18. A database (130) holding an amount of digitized textual information being organized in terms, documents and document corpora, where each document contains at least one term and  
30   each document corpus contains at least one document,  
      each document in a document corpus being associated with concept vector which conceptually classifies the contents of the document on a relatively compact format, and  
      each term in the document corpus being associated with a

term-to-concept vector describing a relationship between the term and each of the concept vectors, characterized in that it is adapted to deliver the term-to-concept vectors to a search engine (115) according to the claim 5 17.

19. A database (130) according to claim 18, characterized in that it comprises an iterative term-to-concept engine adapted to receive fresh digitized textual information added to the database (130) and on basis of this information:  
10 generate concept vectors for any added document, and  
generate a term-to-concept vector describing a relationship between any added term and each of the concept vectors.

20. A server (110) for providing data processing services in respect of digitized textual information, characterized in that it  
15 comprises  
a search engine (115) according to claim 17, and  
a communication interface (112) towards a database (130) according to any one of the claims 18 or 19.

21. A system for providing data processing services in respect  
20 of digitized textual information, characterized in that it comprises  
a server (110) according to claim 20,  
at least one user client (120) adapted to communicate with the server (110), and  
25 a communication link (141; 142) connecting the at least one user client (120) with the server (110).

22. A system according to claim 21, characterized in that an internet (140) accomplishes at least a part of the communication link (141; 142), and the at least one user client (120) comprises



a web browser (121) which in turn provides:

a user input interface (121a) adapted to receive queries (Q) from a user and forward the queries (Q) to the server (110) via the communication link (141), and

- 5 a user output interface (121b) adapted to receive processed textual information (R) from the server (110) via the communication link (142) and present the processed textual information (R) to the user.

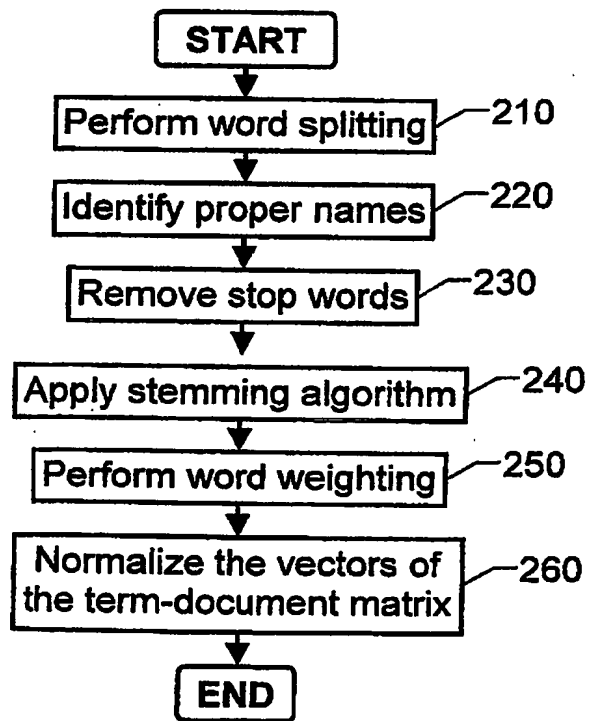
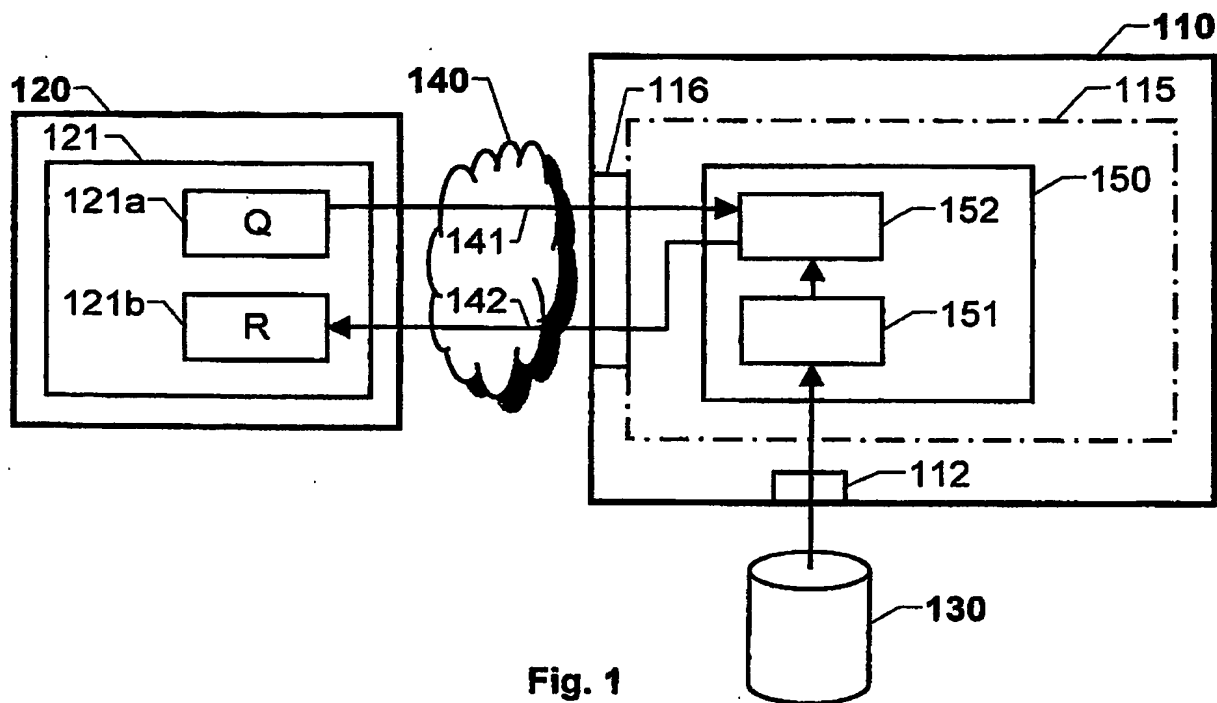
1  
2  
3  
4  
5  
6  
7  
8

**Abstract**

The invention relates to improved solutions for information retrieval, wherein the information is represented by digitized text data. This data is further presumed to be organized in terms  
5 (431 - 438), documents and document corpora, where each document contains at least one term (431 - 438) and each document corpus contains at least one document. Based on a concept vector (420 - 424), which conceptually classifies the contents of each document, a term-to-concept vector is  
10 generated for each term (431 - 438) in the document corpus. The term-to-concept vector describes a relationship between the term (431) and each of the concept vectors (420 - 424). On basis of the term-to-concept vectors for the document corpus, a term-term matrix is generated which describes a term-to-term  
15 relationship between all the terms (431 - 438) in the document corpus. The term-term matrix may then be processed and used for retrieving information from the document corpus, such as the fact that a first term (431) is related to a second term (436).

(Fig. 7b)

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65  
66  
67  
68  
69  
70  
71  
72  
73  
74  
75  
76  
77  
78  
79  
80  
81  
82  
83  
84  
85  
86  
87  
88  
89  
90  
91  
92  
93  
94  
95  
96  
97  
98  
99  
100  
101  
102  
103  
104  
105  
106  
107  
108  
109  
110  
111  
112  
113  
114  
115  
116  
117  
118  
119  
120  
121  
122  
123  
124  
125  
126  
127  
128  
129  
130  
131  
132  
133  
134  
135  
136  
137  
138  
139  
140  
141  
142  
143  
144  
145  
146  
147  
148  
149  
150  
151  
152  
153  
154  
155  
156  
157  
158  
159  
160  
161  
162  
163  
164  
165  
166  
167  
168  
169  
170  
171  
172  
173  
174  
175  
176  
177  
178  
179  
180  
181  
182  
183  
184  
185  
186  
187  
188  
189  
190  
191  
192  
193  
194  
195  
196  
197  
198  
199  
200  
201  
202  
203  
204  
205  
206  
207  
208  
209  
210  
211  
212  
213  
214  
215  
216  
217  
218  
219  
220  
221  
222  
223  
224  
225  
226  
227  
228  
229  
230  
231  
232  
233  
234  
235  
236  
237  
238  
239  
240  
241  
242  
243  
244  
245  
246  
247  
248  
249  
250  
251  
252  
253  
254  
255  
256  
257  
258  
259  
260  
261  
262  
263  
264  
265  
266  
267  
268  
269  
270  
271  
272  
273  
274  
275  
276  
277  
278  
279  
280  
281  
282  
283  
284  
285  
286  
287  
288  
289  
290  
291  
292  
293  
294  
295  
296  
297  
298  
299  
300  
301  
302  
303  
304  
305  
306  
307  
308  
309  
310  
311  
312  
313  
314  
315  
316  
317  
318  
319  
320  
321  
322  
323  
324  
325  
326  
327  
328  
329  
330  
331  
332  
333  
334  
335  
336  
337  
338  
339  
340  
341  
342  
343  
344  
345  
346  
347  
348  
349  
350  
351  
352  
353  
354  
355  
356  
357  
358  
359  
360  
361  
362  
363  
364  
365  
366  
367  
368  
369  
370  
371  
372  
373  
374  
375  
376  
377  
378  
379  
380  
381  
382  
383  
384  
385  
386  
387  
388  
389  
390  
391  
392  
393  
394  
395  
396  
397  
398  
399  
400  
401  
402  
403  
404  
405  
406  
407  
408  
409  
410  
411  
412  
413  
414  
415  
416  
417  
418  
419  
420  
421  
422  
423  
424  
425  
426  
427  
428  
429  
430  
431  
432  
433  
434  
435  
436  
437  
438  
439  
440  
441  
442  
443  
444  
445  
446  
447  
448  
449  
450  
451  
452  
453  
454  
455  
456  
457  
458  
459  
460  
461  
462  
463  
464  
465  
466  
467  
468  
469  
470  
471  
472  
473  
474  
475  
476  
477  
478  
479  
480  
481  
482  
483  
484  
485  
486  
487  
488  
489  
490  
491  
492  
493  
494  
495  
496  
497  
498  
499  
500  
501  
502  
503  
504  
505  
506  
507  
508  
509  
510  
511  
512  
513  
514  
515  
516  
517  
518  
519  
520  
521  
522  
523  
524  
525  
526  
527  
528  
529  
530  
531  
532  
533  
534  
535  
536  
537  
538  
539  
540  
541  
542  
543  
544  
545  
546  
547  
548  
549  
550  
551  
552  
553  
554  
555  
556  
557  
558  
559  
560  
561  
562  
563  
564  
565  
566  
567  
568  
569  
570  
571  
572  
573  
574  
575  
576  
577  
578  
579  
580  
581  
582  
583  
584  
585  
586  
587  
588  
589  
590  
591  
592  
593  
594  
595  
596  
597  
598  
599  
600  
601  
602  
603  
604  
605  
606  
607  
608  
609  
610  
611  
612  
613  
614  
615  
616  
617  
618  
619  
620  
621  
622  
623  
624  
625  
626  
627  
628  
629  
630  
631  
632  
633  
634  
635  
636  
637  
638  
639  
640  
641  
642  
643  
644  
645  
646  
647  
648  
649  
650  
651  
652  
653  
654  
655  
656  
657  
658  
659  
660  
661  
662  
663  
664  
665  
666  
667  
668  
669  
670  
671  
672  
673  
674  
675  
676  
677  
678  
679  
680  
681  
682  
683  
684  
685  
686  
687  
688  
689  
690  
691  
692  
693  
694  
695  
696  
697  
698  
699  
700  
701  
702  
703  
704  
705  
706  
707  
708  
709  
710  
711  
712  
713  
714  
715  
716  
717  
718  
719  
720  
721  
722  
723  
724  
725  
726  
727  
728  
729  
730  
731  
732  
733  
734  
735  
736  
737  
738  
739  
740  
741  
742  
743  
744  
745  
746  
747  
748  
749  
750  
751  
752  
753  
754  
755  
756  
757  
758  
759  
760  
761  
762  
763  
764  
765  
766  
767  
768  
769  
770  
771  
772  
773  
774  
775  
776  
777  
778  
779  
780  
781  
782  
783  
784  
785  
786  
787  
788  
789  
790  
791  
792  
793  
794  
795  
796  
797  
798  
799  
800  
801  
802  
803  
804  
805  
806  
807  
808  
809  
810  
811  
812  
813  
814  
815  
816  
817  
818  
819  
820  
821  
822  
823  
824  
825  
826  
827  
828  
829  
830  
831  
832  
833  
834  
835  
836  
837  
838  
839  
840  
841  
842  
843  
844  
845  
846  
847  
848  
849  
850  
851  
852  
853  
854  
855  
856  
857  
858  
859  
860  
861  
862  
863  
864  
865  
866  
867  
868  
869  
870  
871  
872  
873  
874  
875  
876  
877  
878  
879  
880  
881  
882  
883  
884  
885  
886  
887  
888  
889  
890  
891  
892  
893  
894  
895  
896  
897  
898  
899  
900  
901  
902  
903  
904  
905  
906  
907  
908  
909  
910  
911  
912  
913  
914  
915  
916  
917  
918  
919  
920  
921  
922  
923  
924  
925  
926  
927  
928  
929  
930  
931  
932  
933  
934  
935  
936  
937  
938  
939  
940  
941  
942  
943  
944  
945  
946  
947  
948  
949  
950  
951  
952  
953  
954  
955  
956  
957  
958  
959  
960  
961  
962  
963  
964  
965  
966  
967  
968  
969  
970  
971  
972  
973  
974  
975  
976  
977  
978  
979  
980  
981  
982  
983  
984  
985  
986  
987  
988  
989  
990  
991  
992  
993  
994  
995  
996  
997  
998  
999  
1000  
1001  
1002  
1003  
1004  
1005  
1006  
1007  
1008  
1009  
1010  
1011  
1012  
1013  
1014  
1015  
1016  
1017  
1018  
1019  
1020  
1021  
1022  
1023  
1024  
1025  
1026  
1027  
1028  
1029  
1030  
1031  
1032  
1033  
1034  
1035  
1036  
1037  
1038  
1039  
1040  
1041  
1042  
1043  
1044  
1045  
1046  
1047  
1048  
1049  
1050  
1051  
1052  
1053  
1054  
1055  
1056  
1057  
1058  
1059  
1060  
1061  
1062  
1063  
1064  
1065  
1066  
1067  
1068  
1069  
1070  
1071  
1072  
1073  
1074  
1075  
1076  
1077  
1078  
1079  
1080  
1081  
1082  
1083  
1084  
1085  
1086  
1087  
1088  
1089  
1090  
1091  
1092  
1093  
1094  
1095  
1096  
1097  
1098  
1099  
1100  
1101  
1102  
1103  
1104  
1105  
1106  
1107  
1108  
1109  
1110  
1111  
1112  
1113  
1114  
1115  
1116  
1117  
1118  
1119  
1120  
1121  
1122  
1123  
1124  
1125  
1126  
1127  
1128  
1129  
1130  
1131  
1132  
1133  
1134  
1135  
1136  
1137  
1138  
1139  
1140  
1141  
1142  
1143  
1144  
1145  
1146  
1147  
1148  
1149  
1150  
1151  
1152  
1153  
1154  
1155  
1156  
1157  
1158  
1159  
1160  
1161  
1162  
1163  
1164  
1165  
1166  
1167  
1168  
1169  
1170  
1171  
1172  
1173  
1174  
1175  
1176  
1177  
1178  
1179  
1180  
1181  
1182  
1183  
1184  
1185  
1186  
1187  
1188  
1189  
1190  
1191  
1192  
1193  
1194  
1195  
1196  
1197  
1198  
1199  
1200  
1201  
1202  
1203  
1204  
1205  
1206  
1207  
1208  
1209  
1210  
1211  
1212  
1213  
1214  
1215  
1216  
1217  
1218  
1219  
1220  
1221  
1222  
1223  
1224  
1225  
1226  
1227  
1228  
1229  
1230  
1231  
1232  
1233  
1234  
1235  
1236  
1237  
1238  
1239  
1240  
1241  
1242  
1243  
1244  
1245  
1246  
1247  
1248  
1249  
1250  
1251  
1252  
1253  
1254  
1255  
1256  
1257  
1258  
1259  
1260  
1261  
1262  
1263  
1264  
1265  
1266  
1267  
1268  
1269  
1270  
1271  
1272  
1273  
1274  
1275  
1276  
1277  
1278  
1279  
1280  
1281  
1282  
1283  
1284  
1285  
1286  
1287  
1288  
1289  
1290  
1291  
1292  
1293  
1294  
1295  
1296  
1297  
1298  
1299  
1300  
1301  
1302  
1303  
1304  
1305  
1306  
1307  
1308  
1309  
1310  
1311  
1312  
1313  
1314  
1315  
1316  
1317  
1318  
1319  
1320  
1321  
1322  
1323  
1324  
1325  
1326  
1327  
1328  
1329  
1330  
1331  
1332  
1333  
1334  
1335  
1336  
1337  
1338  
1339  
1340  
1341  
1342  
1343  
1344  
1345  
1346  
1347  
1348  
1349  
1350  
1351  
1352  
1353  
1354  
1355  
1356  
1357  
1358  
1359  
1360  
1361  
1362  
1363  
1364  
1365  
1366  
1367  
1368  
1369  
1370  
1371  
1372  
1373  
1374  
1375  
1376  
1377  
1378  
1379  
1380  
1381  
1382  
1383  
1384  
1385  
1386  
1387  
1388  
1389  
1390  
1391  
1392  
1393  
1394  
1395  
1396  
1397  
1398  
1399  
1400  
1401  
1402  
1403  
1404  
1405  
1406  
1407  
1408  
1409  
1410  
1411  
1412  
1413  
1414  
1415  
1416  
1417  
1418  
1419  
1420  
1421  
1422  
1423  
1424  
1425  
1426  
1427  
1428  
1429  
1430  
1431  
1432  
1433  
1434  
1435  
1436  
1437  
1438  
1439  
1440  
1441  
1442  
1443  
1444  
1445  
1446  
1447  
1448  
1449  
1450  
1451  
1452  
1453  
1454  
1455  
1456  
1457  
1458  
1459  
1460  
1461  
1462  
1463  
1464  
1465  
1466  
1467  
1468  
1469  
1470  
1471  
1472  
1473  
1474  
1475  
1476  
1477  
1478  
1479  
1480  
1481  
1482  
1483  
1484  
1485  
1486  
1487  
1488  
1489  
1490  
1491  
1492  
1493  
1494  
1495  
1496  
1497  
1498  
1499  
1500  
1501  
1502  
1503  
1504  
1505  
1506  
1507  
1508  
1509  
1510  
1511  
1512  
1513  
1514  
1515  
1516  
1517  
1518  
1519  
1520  
1521  
1522  
1523  
1524  
1525  
1526  
1527  
1528  
1529  
1530  
1531  
1532  
1533  
1534  
1535  
1536  
1537  
1538  
1539  
1540  
1541  
1542  
1543  
1544  
1545  
1546  
1547  
1548  
1549  
1550  
1551  
1552  
1553  
1554  
1555  
1556  
1557  
1558  
1559  
1560  
1561  
1562  
1563  
1564  
1565  
1566  
1567  
1568  
1569  
1570  
1571  
1572  
1573  
1574  
1575  
1576  
1577  
1578  
1579  
1580  
1581  
1582  
1583  
1584  
1585  
1586  
1587  
1588  
1589  
1590  
1591  
1592  
1593  
1594  
1595  
1596  
1597  
1598  
1599  
1600  
1601  
1602  
1603  
1604  
1605  
1606  
1607  
1608  
1609  
1610  
1611  
1612  
1613  
1614  
1615  
1616  
1617  
1618  
1619  
1620  
1621  
1622  
1623  
1624  
1625  
1626  
1627  
1628  
1629  
1630  
1631  
1632  
1633  
1634  
1635  
1636  
1637  
1638  
1639  
1640  
1641  
1642  
1643  
1644  
1645  
1646  
1647  
1648  
1649  
1650  
1651  
1652  
1653  
1654  
1655  
1656  
1657  
1658  
1659  
1660  
1661  
1662  
1663  
1664  
1665  
1666  
1667  
1668  
1669  
1670  
1671  
1672  
1673  
1674  
1675  
1676  
1677  
1678  
1679  
1680  
1681  
1682  
1683  
1684  
1685  
1686  
1687  
1688  
1689  
1690  
1691  
1692  
1693  
1694  
1695  
1696  
1697  
1698  
1699  
1700  
1701  
1702  
1703  
1704  
1705  
1706  
1707  
1708  
1709  
1710  
1711  
1712  
1713  
1714  
1715  
1716  
1717  
1718  
1719  
1720  
1721  
1722  
1723  
1724  
1725  
1726  
1727  
1728  
1729  
1730  
1731  
1732  
1733  
1734  
1735  
1736  
1737  
1738  
1739  
1740  
1741  
1742  
1743  
1744  
1745  
1746  
1747  
1748  
1749  
1750  
1751  
1752  
1753  
1754  
1755  
1756  
1757  
1758  
1759  
1760  
1761  
1762  
1763  
1764  
1765  
1766  
1767  
1768  
1769  
1770  
1771  
1772  
1773  
1774  
1775  
1776  
1777  
1778  
1779  
1780  
1781  
1782  
1783  
1784  
1785  
1786  
1787  
1788  
1789  
1790  
1791  
1792  
1793  
1794  
1795  
1796  
1797  
1798  
1799  
1800  
1801  
1802  
1803  
1804  
1805  
1806  
1807  
1808  
1809  
1810  
1811  
1812  
1813  
1814  
1815  
1816  
1817  
1818  
1819  
1820  
1821  
1822  
1823  
1824  
1825  
1826  
1827  
1828  
1829  
1830  
1831  
1832  
1833  
1834  
1835  
1836  
1837  
1838  
1839  
1840  
1841  
1842  
1843  
1844  
1845  
1846  
1847  
1848  
1849  
1850  
1851  
1852  
1853  
1854  
1855  
1856  
1857  
1858  
1859  
1860  
1861  
1862  
1863  
1864  
1865  
1866  
1867  
1868  
1869  
1870  
1871  
1872  
1873  
1874  
1875  
1876  
1877  
1878  
1879  
1880  
1881  
1882  
1883  
1884  
1885  
1886  
1887  
1888  
1889  
1890  
1891  
1892  
1893  
1894  
1895  
1896  
1897  
1898  
1899  
1900  
1901  
1902  
1903  
1904  
1905  
1906  
1907  
1908  
1909  
1910  
1911  
1912  
1913  
1914  
1915  
1916  
1917  
1918  
1919  
1920  
1921  
1922  
1923  
1924  
1925  
1926  
1927  
1928  
1929  
1930  
1931  
1932  
1933  
1934  
1935  
1936  
1937  
1938  
1939  
1940  
1941  
1942  
1943  
1944  
1945  
1946  
1947  
1948  
1949  
1950  
1951  
1952  
1953  
1954  
1955  
1956  
1957  
1958  
1959  
1960  
1961  
1962  
1963  
1964  
1965  
1966  
1967  
1968  
1969  
1970  
1971  
1972  
1973  
1974  
1975  
1976  
1977  
1978  
1979  
1980  
1981  
1982  
1983  
1984  
1985  
1986  
1987  
1988  
1989  
1990  
1991  
1992  
1993  
1994  
1995  
1996  
1997  
1998  
1999  
2000  
2001  
2002  
2003  
2004  
2005  
2006  
2007  
2008  
2009  
2010  
2011  
2012  
2013  
2014  
2015  
2016  
2017  
2018  
2019  
2020  
2021  
2022  
2023  
2024  
2025  
2026  
2027  
2028  
2029  
2030  
2031  
2032  
2033  
2034  
2035  
2036  
2037  
2038  
2039  
2040  
2041  
2042  
2043  
2044  
2045  
2046  
2047  
2048  
2049  
2050  
2051  
2052  
2053  
2054  
2055  
2056  
2057  
2058  
2059  
2060  
2061  
2062  
2063  
2064  
2065  
2066  
2067  
2068  
2069  
2070  
2071  
2072  
2073  
2074  
2075  
2076  
2077  
2078  
2079  
2080  
2081  
2082  
2083  
2084  
2085  
2086  
2087  
2088  
2089  
2090  
2091  
2092  
2093  
2094  
2095  
2096  
2097  
2098  
2099  
2100  
2101  
2102  
2103  
2104  
2105  
2106  
2107  
2108  
2109  
2110  
2111  
2112  
2113  
2114  
2115  
2116  
2117  
2118  
2119  
2120  
2121  
2122  
2123  
2124  
2125  
2126  
2127  
2128  
2129  
2130  
2131  
2132  
2133  
2134  
2135  
2136  
2137  
2138  
2139  
2140  
2141  
2142  
2143  
2144  
2145  
2146  
2147  
2148  
2149  
2150  
2151  
2152  
2153  
2154  
2155  
2156  
2157  
2158  
2159  
2160  
2161  
2162  
2163  
2164  
2165  
2166  
2167  
2168  
2169  
2170  
2171  
21



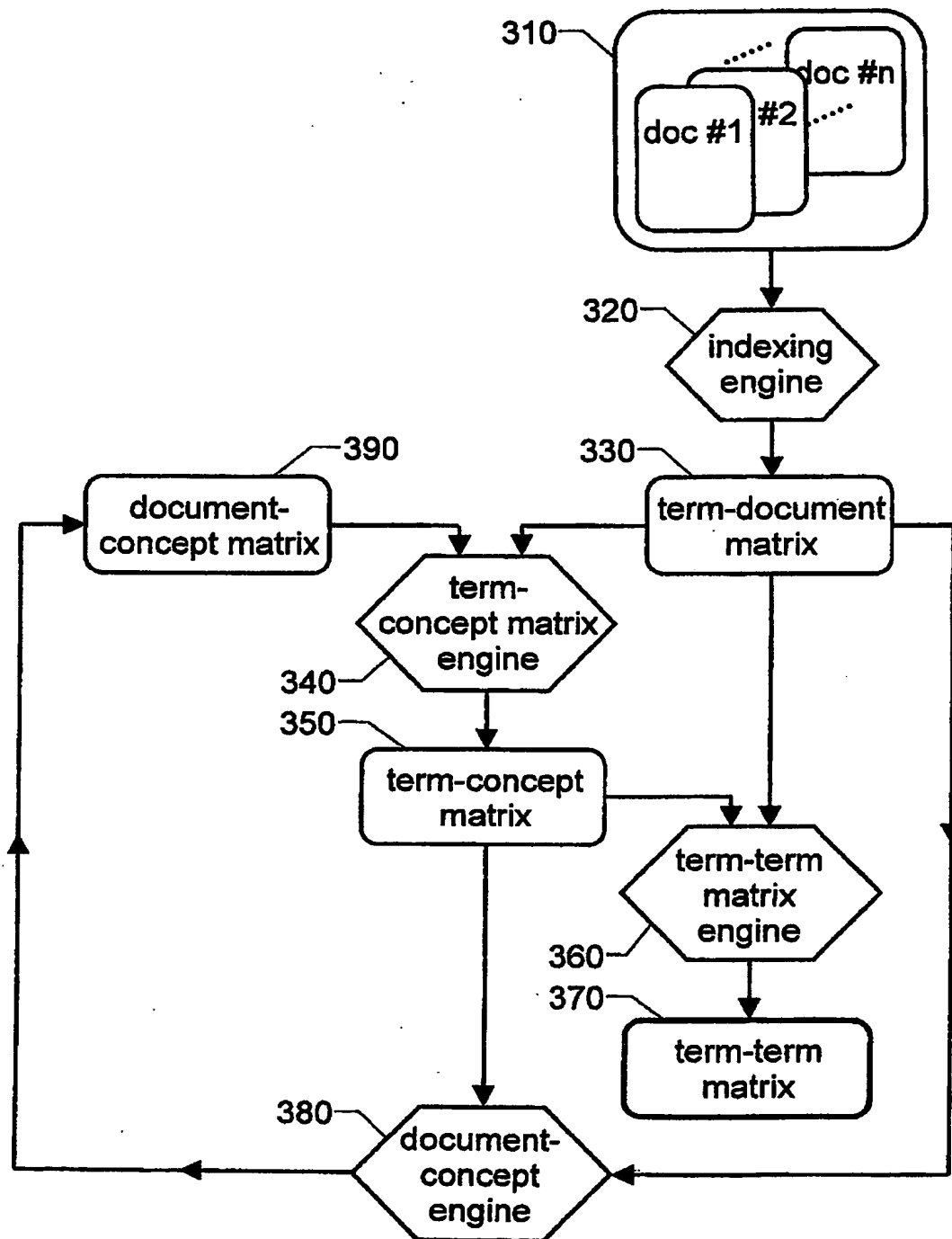


Fig. 3

3/9

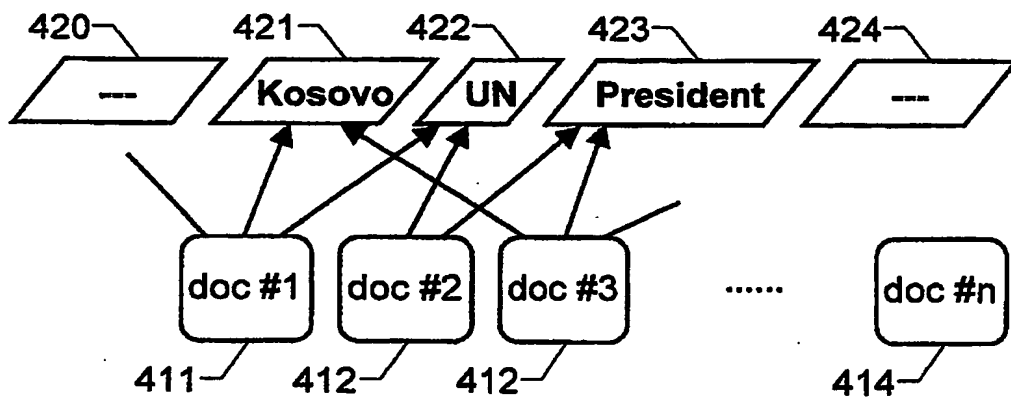


Fig. 4a

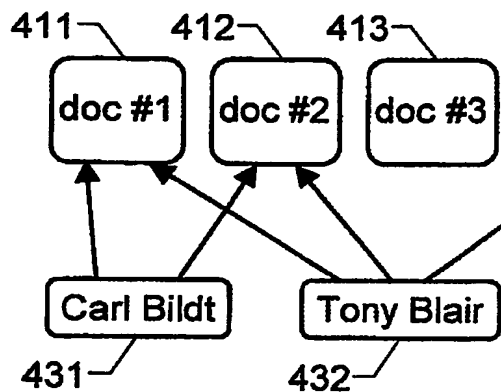
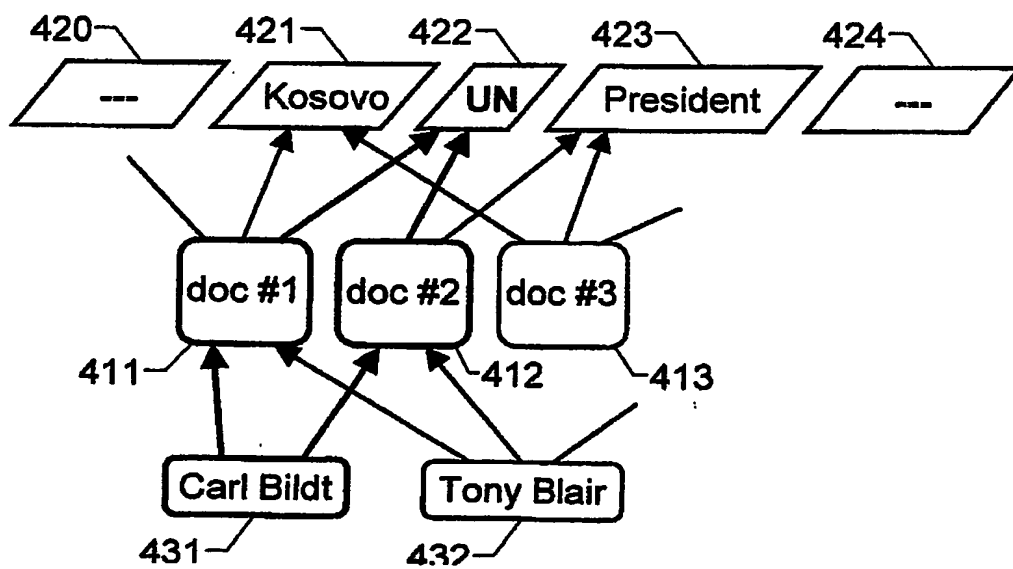


Fig. 4b



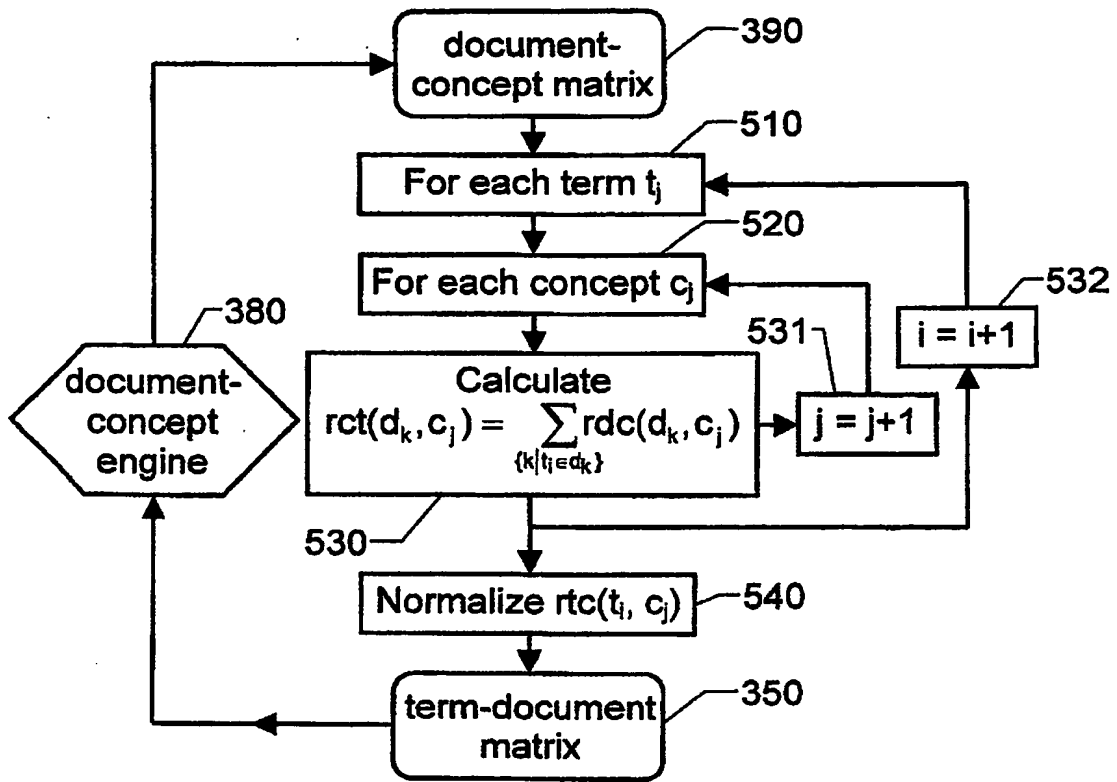


Fig. 5

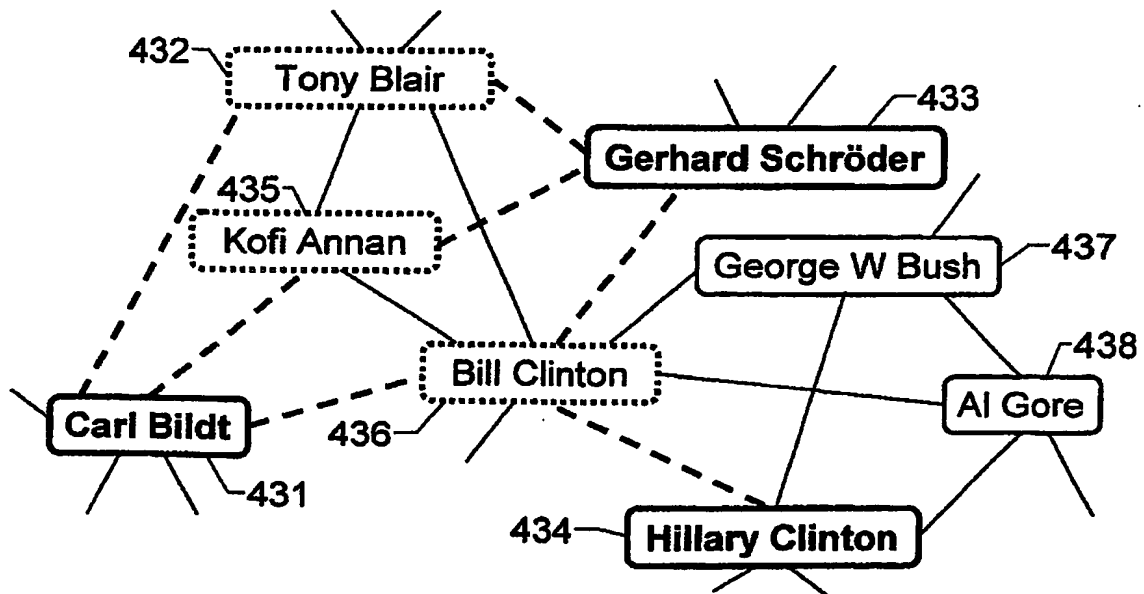


Fig. 11

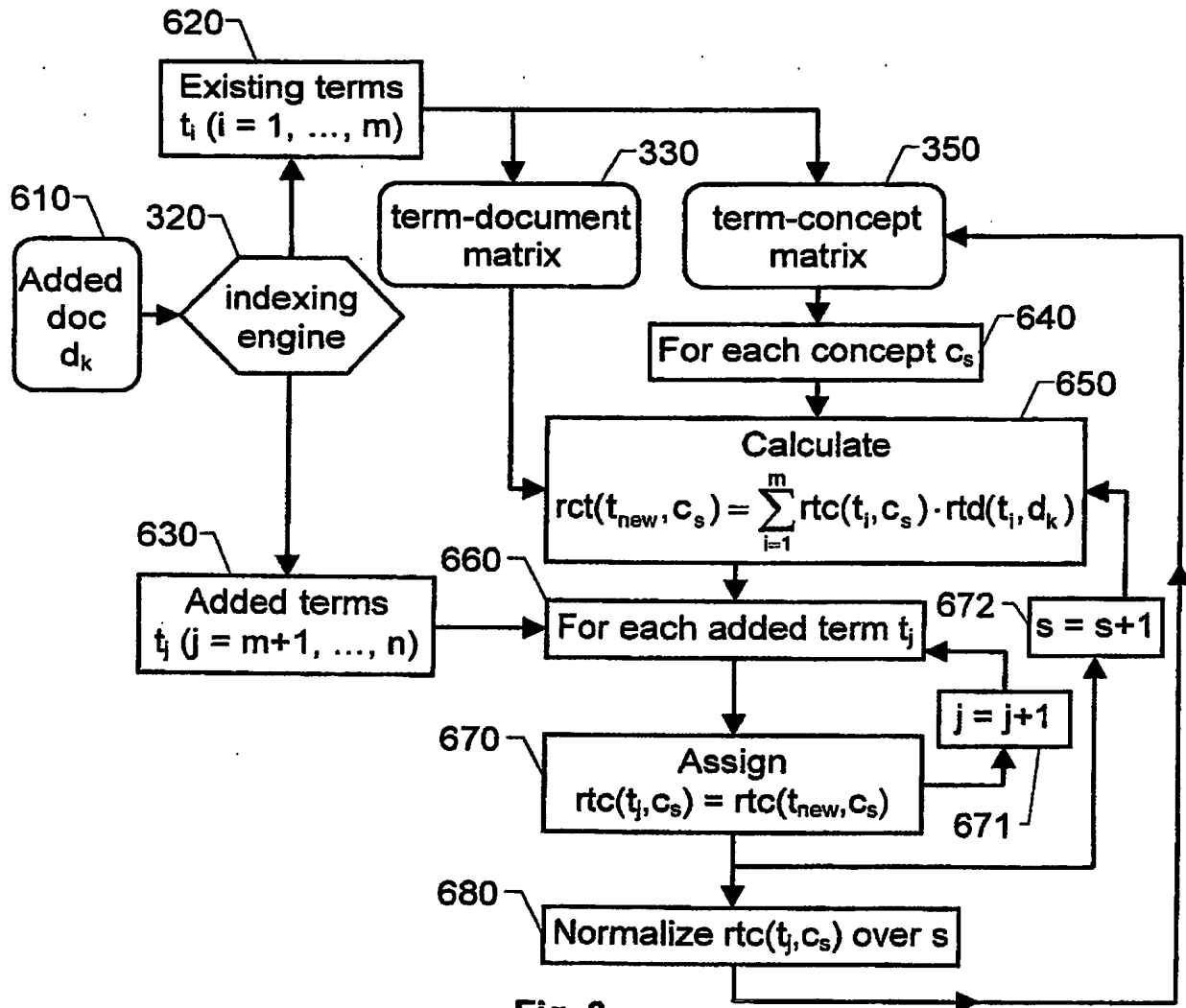


Fig. 6

6/9

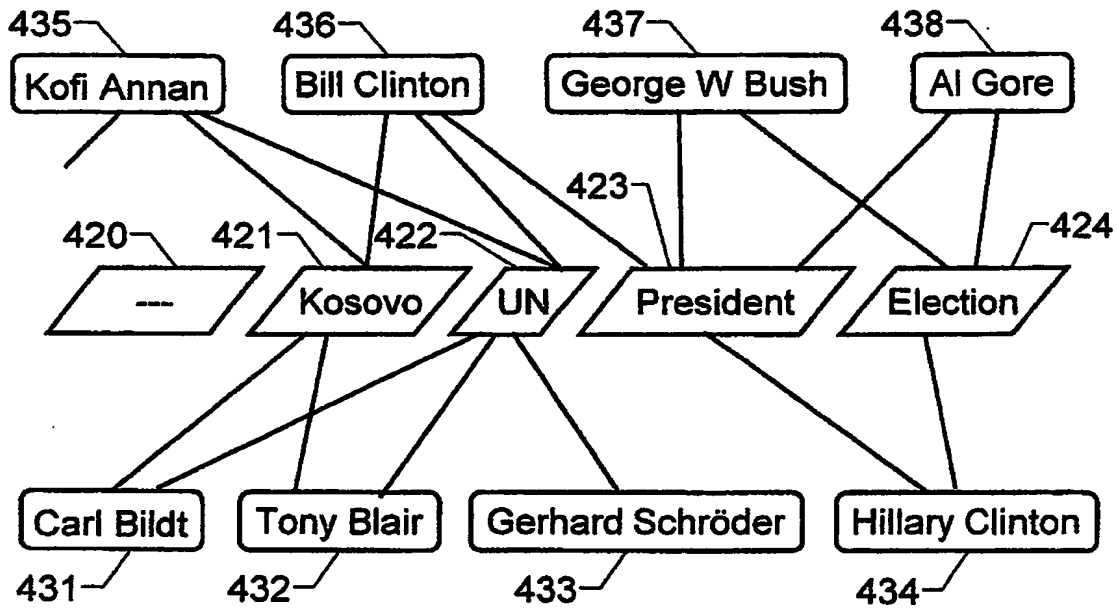


Fig. 7a

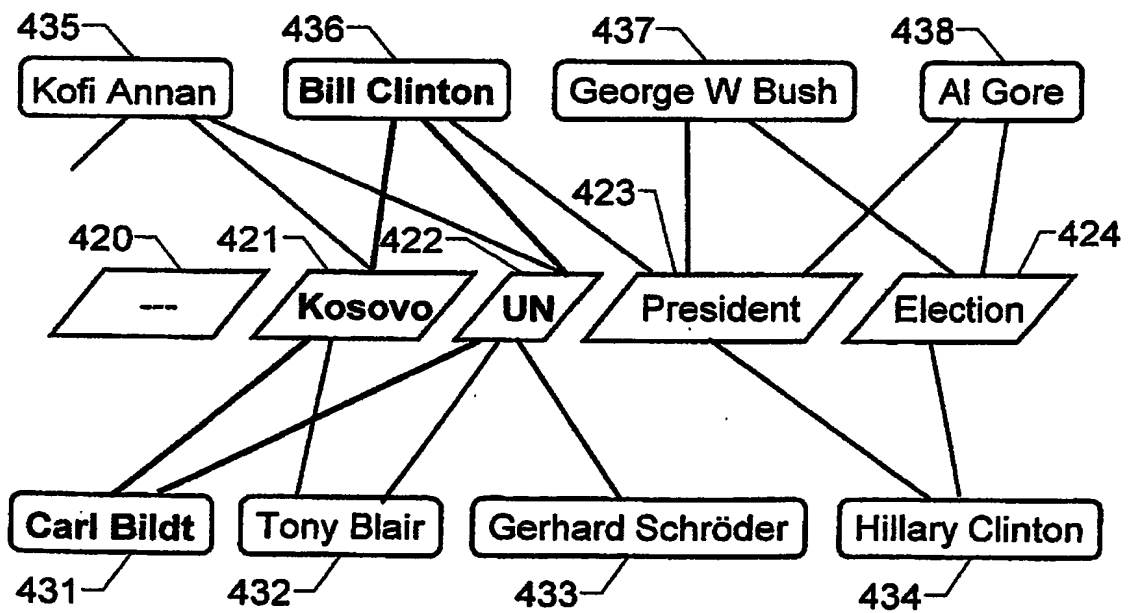


Fig. 7b



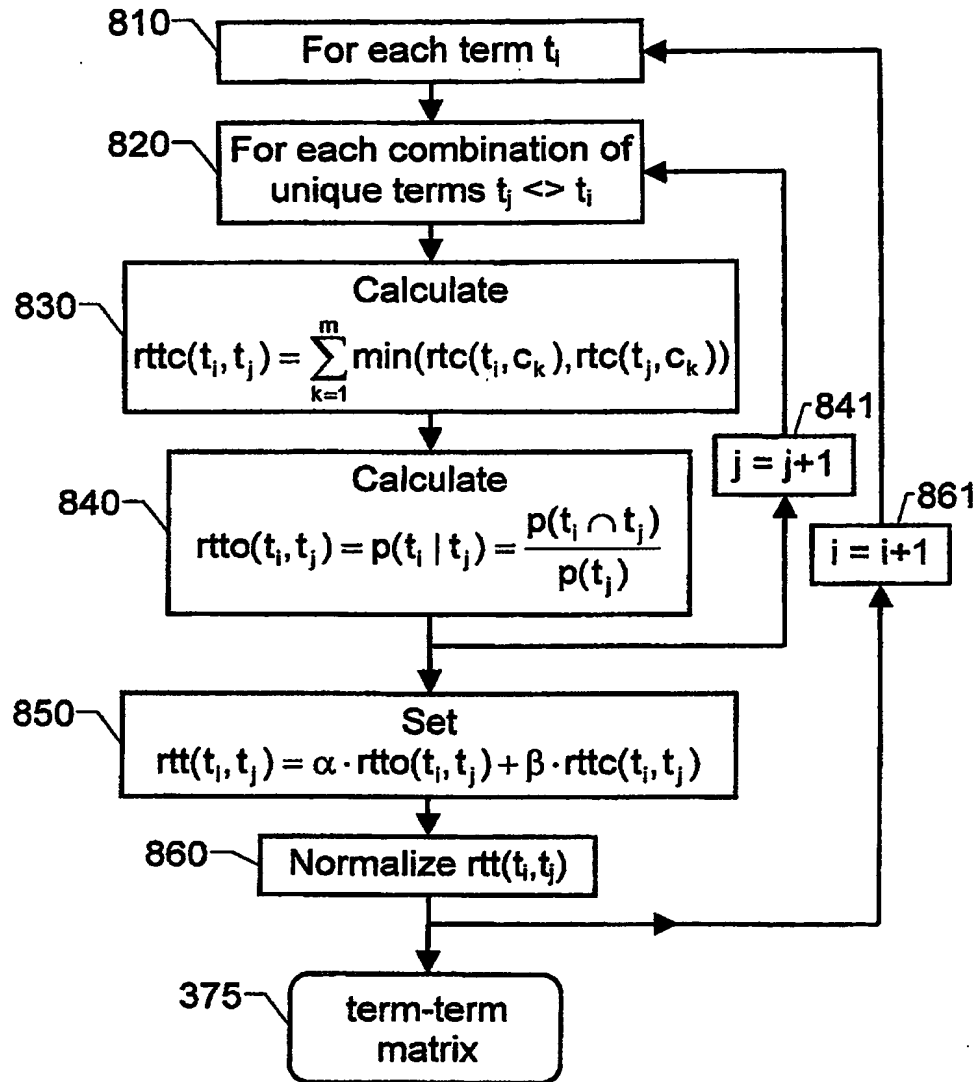
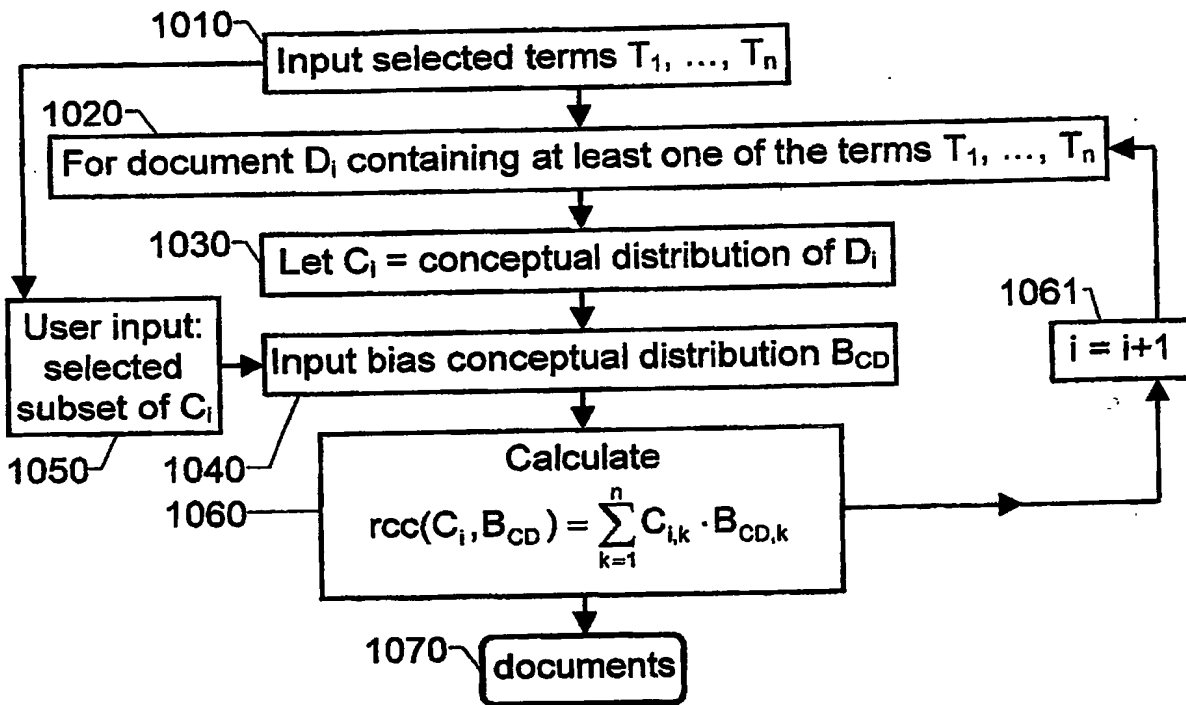
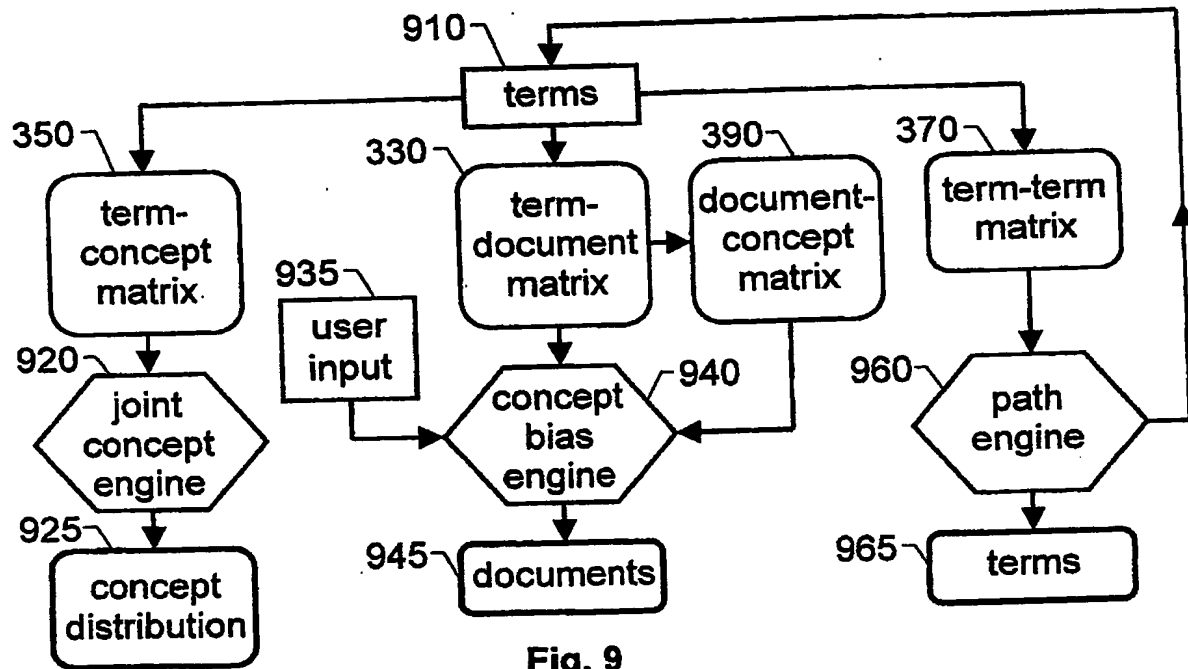


Fig. 8

8/9



9/9

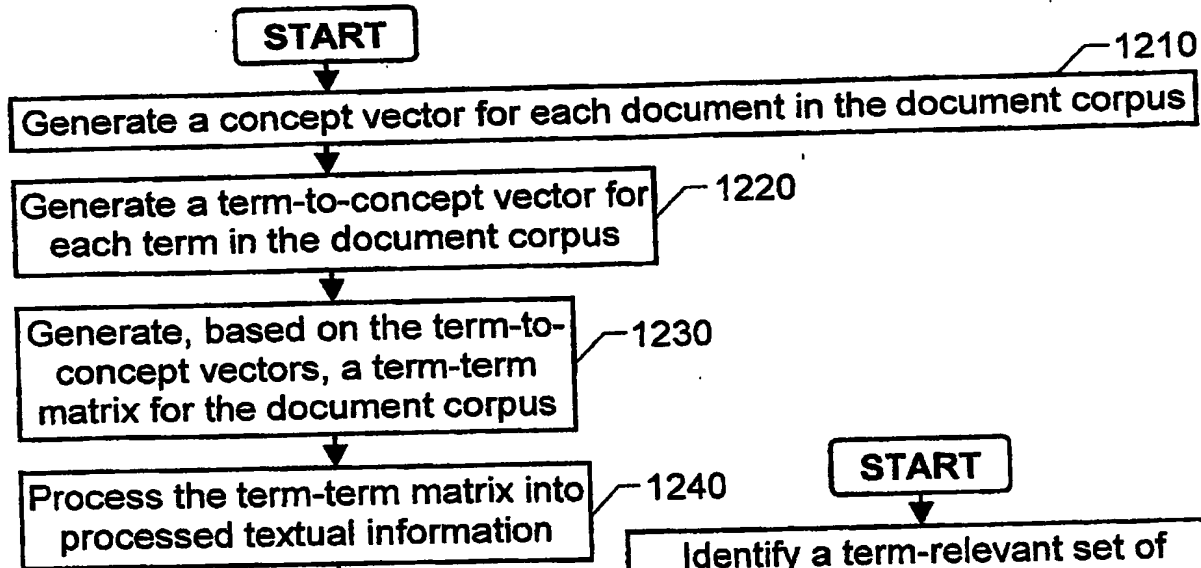


Fig. 12

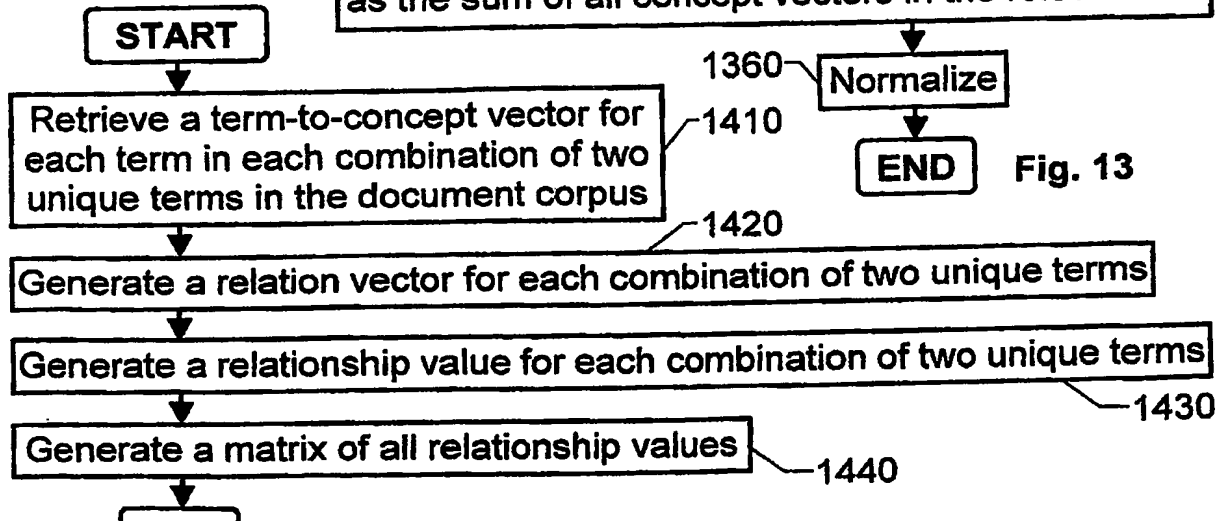
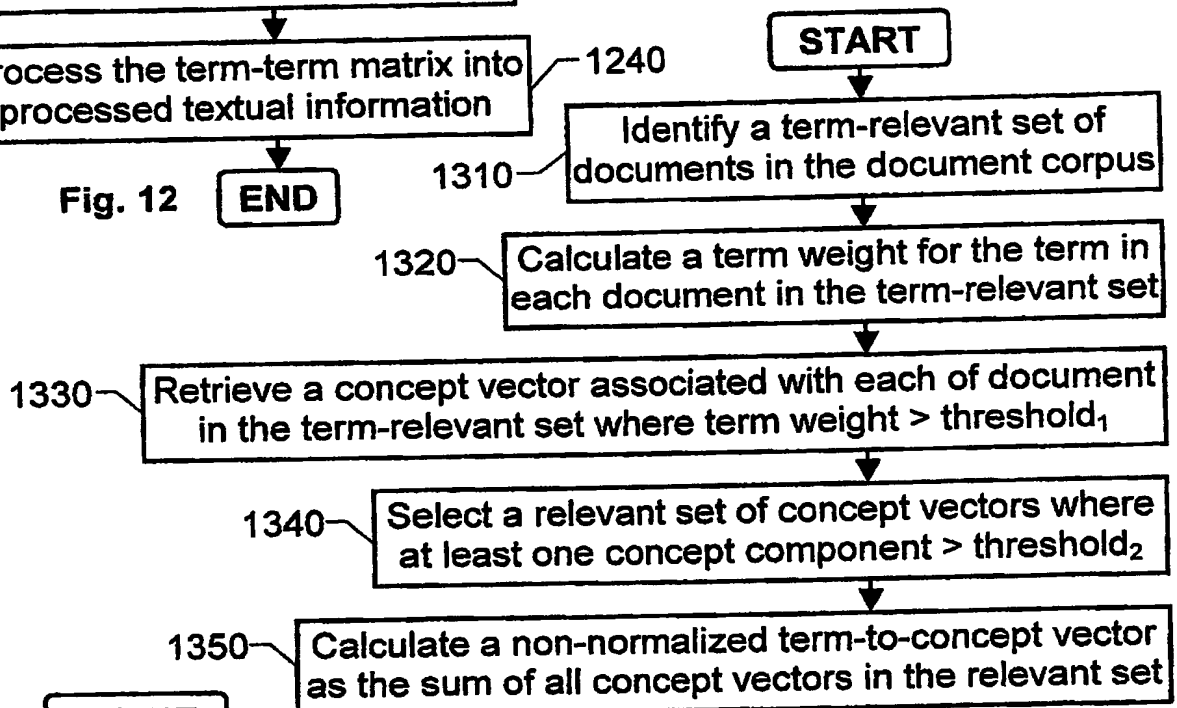


Fig. 13

**This Page is Inserted by IFW Indexing and Scanning  
Operations and is not part of the Official Record**

**BEST AVAILABLE IMAGES**

Defective images within this document are accurate representations of the original documents submitted by the applicant.

Defects in the images include but are not limited to the items checked:

- ☐ **BLACK BORDERS**
- ☐ **IMAGE CUT OFF AT TOP, BOTTOM OR SIDES**
- ☐ **FADED TEXT OR DRAWING**
- ☒ **BLURRED OR ILLEGIBLE TEXT OR DRAWING**
- ☐ **SKEWED/SLANTED IMAGES**
- ☐ **COLOR OR BLACK AND WHITE PHOTOGRAPHS**
- ☐ **GRAY SCALE DOCUMENTS**
- ☒ **LINES OR MARKS ON ORIGINAL DOCUMENT**
- ☐ **REFERENCE(S) OR EXHIBIT(S) SUBMITTED ARE POOR QUALITY**
- ☐ **OTHER:** \_\_\_\_\_

**IMAGES ARE BEST AVAILABLE COPY.**

**As rescanning these documents will not correct the image problems checked, please do not report these problems to the IFW Image Problem Mailbox.**